GENEALOGIES OF TWO LINKED NEUTRAL LOCI AFTER A SELECTIVE SWEEP IN A LARGE POPULATION OF STOCHASTICALLY VARYING SIZE

REBEKKA BRINK-SPALINK AND CHARLINE SMADI

Abstract. We study the impact of a hard selective sweep on the genealogy of partially linked neutral loci in the vicinity of the positively selected allele. We consider a sexual population of stochastically varying size and, focusing on two neighboring loci, derive an approximate formula for the neutral genealogy of a sample of individuals taken at the end of the sweep. Individuals are characterized by ecological parameters depending on their genetic type and governing their growth rate and interactions with other individuals (competition). As a consequence, the "fitness" of an individual depends on the population state and is not an intrinsic characteristic of individuals. We provide a deep insight into the dynamics of the mutant and wild type populations during the different stages of a selective sweep.

Introduction

We study the hitchhiking effect of a beneficial mutation in a sexual haploid population of stochastically varying size. We assume that a mutation occurs in one individual of a monomorphic population and that individuals carrying the new allele a are better adapted to the current environment and spread in the population. We suppose that the mutant allele a eventually replaces the resident one, A, and study the influence of this fixation on the neutral gene genealogy of a sample taken at the end of the selective sweep. That is, in each sampled individual we consider the same set of partially linked loci including the locus where the advantageous mutation occurred. We then trace back the ancestral lineages of all loci in the sample until the beginning of the sweep and update the genetic relationships whenever a coalescence or a recombination changes the ancestry of one or several loci. Our main result is the derivation of a sampling formula for the ancestral partition of two neutral loci situated in the vicinity of the selected allele.

The first studies of hitchhiking, initiated by Maynard Smith and Haigh [20], have modeled the mutant population size as the solution of a deterministic logistic equation [16, 13, 22, 21]. Barton [1] was the first to point out the importance of the stochasticity of the mutant population size. Following this paper, a series of works took into account this randomness during the sweep. In [9, 18] Schweinsberg and Durrett based their analysis on a Moran model with selection and recombination, while Etheridge and coauthors [10] worked with the diffusion limit of such discrete population models. Then Brink-Spalink [2], Pfaffelhuber and Studeny [17] and Leocard [14] extended the respective findings of these two approaches for the ancestry of one neutral locus to the two-locus (resp. multiple-locus) case.

However, in all these models, the population size was constant and each individual had a "fitness" only dependent on its type and not on the population state. The fundamental idea of Darwin is that the individual traits have an influence on the interactions between individuals, which in turn generate selection on the different traits. In this paper we aim at modeling precisely these interactions by extending the model introduced in [19] where the author considered only one neutral locus. Such an eco-evolutionary approach has been introduced by Metz and coauthors [15] and has been made rigorous in the seminal paper of Fournier and Méléard [12]. Then it was further developed by Champagnat, Méléard and coauthors (see [3, 5, 4] and references therein) for the haploid asexual case and by Collet, Méléard and Metz [6] and Coron and coauthors [7] for the diploid sexual case.

Institute for Mathematical Stochastics, Georg-August-Universität Göttingen, Goldschmidtstr. 7, 37077 Göttingen, Germany

Irstea, UR LISC, Laboratoire d'Ingénierie des Systèmes Complexes, 9 avenue Blaise Pascal-CS 20085, 63178 Aubière, France and Department of Statistics, University of Oxford, 1 South Parks Road, Oxford OX1 3TG, UK

E-mail addresses: rbrink@math.uni-goettingen.de , charline.smadi@polytechnique.edu.

2010 *Mathematics Subject Classification.* 92D25, 60J80, 60J27, 92D15, 60F15.

Key words and phrases. Birth and death process; Coalescent; Coupling; Eco-evolution; Genetic hitch-hiking; Selective sweep.

The population dynamics, described in Section 1, is a multitype birth and death Markov process with competition. We represent the carrying capacity of the underlying environment by a scaling parameter $K \in \mathbb{N}$ and state results in the limit for large K. In [3] it was shown that such kind of invasion processes can be divided into three phases (see Figure 2): an initial phase in which the fraction of a-individuals does not exceed a fixed value $\varepsilon > 0$ and where the dynamics of the wild type population is nearly undisturbed by the invading type. A second phase where both types account for a non-negligible percentage of the population and where the dynamics of the population can be well approximated by a deterministic competitive Lotka-Volterra system. And finally a third phase where the roles of the types are interchanged and the wild type population is near extinction. The durations of the first and third phases of the selective sweep are of order $\log K$ whereas the second phase only lasts an amount of time of order 1. This three phases decomposition is commonly encountered in population genetics models and dates back to [16].

In Section 3 we precisely describe these three phases and introduce two couplings of the population process, key tools to study the dynamics of the A- and a-populations. Section 4 is devoted to the proofs of the main theorems on the ancestral partition of the two neutral alleles. Sections 5 to 7 are dedicated to the proofs of auxiliary statements. In Section 2 we compare our findings with previous results. Finally, we state technical results needed in the proofs in the Appendix.

1. Model and results

We consider a three locus model: one locus under selection, SL, with alleles in $\mathcal{A} := \{A, a\}$ and two neighboring neutral loci $N1$ and $N2$ with alleles in the finite sets $\mathcal{B}$ and $\mathcal{C}$ respectively. We denote by $\mathcal{E} = \mathcal{A} \times \mathcal{B} \times \mathcal{C}$ the type space. Two geometric alignments are possible: either the two neutral loci are adjacent (geometry $SL-N1-N2$), or they are separated by the selected locus (geometry $N1-SL-N2$). We introduce the model and notations for the adjacent geometry, their analogs for the separated one can be deduced in a straightforward manner.

Whenever a reproduction event takes place, recombinations between SL and $N1$ or between $N1$ and $N2$ occur independently with probabilities r_1 and r_2, respectively. These probabilities depend on the parameter K, representing the environment's carrying capacity, but for the purpose of readability we do not indicate this dependence. We assume a regime of weak recombination:

$$\limsup_{K\to\infty}\ r_j \log K < \infty,\ \ j = 1, 2. \tag{1.1}$$

This is motivated by Theorem 2 in [19] which states that this is the good scale to observe a signature on the neutral allele distribution. If the recombination probabilities are larger (neutral loci more distant from the selected locus), there are many recombinations and the sweep does not modify the neutral diversity at these sites. Recombinations may lead to a mixing of the parental genetic material in the newborn, and hence, parents with types $\alpha\beta\gamma$ and $\alpha'\beta'\gamma'$ in $\mathcal{E}$ can generate the following offspring:

possible genotype	event	probability
$\alpha\beta\gamma, \alpha'\beta'\gamma'$	no recombination	$(1-r_1)(1-r_2)$
$\alpha\beta'\gamma', \alpha'\beta\gamma$	one recombination between SL and $N1$	$r_1(1-r_2)$
$\alpha\beta\gamma', \alpha'\beta'\gamma$	one recombination between $N1$ and $N2$	$(1-r_1)r_2$
$\alpha\beta'\gamma, \alpha'\beta\gamma'$	two recombinations	r_1r_2

We will see in the sequel that the probability to witness a birth event with two simultaneous recombinations in the neutral genealogy of a uniformly chosen individual is very small.

As we assume the loci $N1$ and $N2$ to be neutral, the ecological parameters of an individual only depend on the allele α at the locus under selection. Let us denote by f_α the fertility of an individual with type α. In the spirit of [6], such an individual gives birth at rate f_α (female role), and has a probability proportional to f_α to be chosen as the father in a given birth event (male role). Denoting the complementary type of the allele α by $\bar{\alpha}$ we get the following result for the birth rate of individuals of

type $\alpha\beta\gamma \in \mathcal{E}$:

$$(1.2)\quad b^K_{\alpha\beta\gamma}(n) = (1-r_1)(1-r_2)f_\alpha n_{\alpha\beta\gamma} + r_1(1-r_2)f_\alpha n_\alpha \frac{f_\alpha n_{\alpha\beta\gamma} + f_{\bar\alpha} n_{\bar\alpha\beta\gamma}}{f_a n_a + f_A n_A} +$$
$$(1-r_1)r_2 f_\alpha \frac{\sum_{(\beta',\gamma')\in(\mathcal{B},\mathcal{C})} n_{\alpha\beta\gamma'}(f_\alpha n_{\alpha\beta'\gamma} + f_{\bar\alpha} n_{\bar\alpha\beta'\gamma})}{f_a n_a + f_A n_A} + r_1 r_2 f_\alpha \frac{\sum_{(\beta',\gamma')\in(\mathcal{B},\mathcal{C})} n_{\alpha\beta'\gamma}(f_\alpha n_{\alpha\beta\gamma'} + f_{\bar\alpha} n_{\bar\alpha\beta\gamma'})}{f_a n_a + f_A n_A},$$

where $n_{\alpha\beta\gamma}$ (resp. n_α) denotes the current number of $\alpha\beta\gamma$-individuals (resp. α-individuals) and $n = (n_{\alpha\beta\gamma}, (\alpha,\beta,\gamma) \in \mathcal{E})$ is the current state of the population. An α-individual can die either from a natural death (rate D_α), or from type-dependent competition: the parameter $C_{\alpha,\alpha'}$ models the impact an individual of type α' has on an individual of type α, where $(\alpha,\alpha') \in \mathcal{A}^2$. The strength of the competition also depends on the carrying capacity K. This results in the total death rate of individuals carrying the alleles $\alpha\beta\gamma \in \mathcal{E}$:

$$(1.3)\qquad d^K_{\alpha\beta\gamma}(n) = \Big(D_\alpha + \frac{C_{\alpha,A}}{K} n_A + \frac{C_{\alpha,a}}{K} n_a\Big) n_{\alpha\beta\gamma}.$$

Hence the population process

$$N^K = (N^K(t), t \geq 0) = \Big((N^K_{\alpha\beta\gamma}(t))_{(\alpha,\beta,\gamma)\in\mathcal{E}}, t \geq 0\Big),$$

where $N^K_{\alpha\beta\gamma}(t)$ denotes the number of $\alpha\beta\gamma$-individuals at time t, is a multitype birth and death process with rates given in (1.2) and (1.3). We will often work with the trait population process $((N^K_A(t), N^K_a(t)), t \geq 0)$, where $N^K_\alpha(t)$ denotes the number of α-individuals at time t. This is also a birth and death process with birth and death rates given by:

$$(1.4)\qquad b^K_\alpha(n) = \sum_{(\beta,\gamma)\in\mathcal{B}\times\mathcal{C}} b^K_{\alpha\beta\gamma}(n) = f_\alpha n_\alpha$$
$$d^K_\alpha(n) = \sum_{(\beta,\gamma)\in\mathcal{B}\times\mathcal{C}} d^K_{\alpha\beta\gamma}(n) = \Big(D_\alpha + \frac{C_{\alpha,A}}{K} n_A + \frac{C_{\alpha,a}}{K} n_a\Big) n_\alpha.$$

As a quantity summarizing the advantage or disadvantage a mutant with allele type α has in an $\bar\alpha$-population at equilibrium, we introduce the so-called invasion fitness $S_{\alpha\bar\alpha}$ through

$$(1.5)\qquad S_{\alpha\bar\alpha} := f_\alpha - D_\alpha - C_{\alpha,\bar\alpha}\bar n_{\bar\alpha},$$

where the equilibrium density $\bar n_\alpha$ is defined by

$$(1.6)\qquad \bar n_\alpha := \frac{f_\alpha - D_\alpha}{C_{\alpha,\alpha}}.$$

The role of the invasion fitness $S_{\alpha\bar\alpha}$ and the definition of the equilibrium density $\bar n_\alpha$ follow from the properties of the two-dimensional competitive Lotka-Volterra system:

$$(1.7)\qquad \dot n^{(z)}_\alpha = (f_\alpha - D_\alpha - C_{\alpha,A} n^{(z)}_A - C_{\alpha,a} n^{(z)}_a) n^{(z)}_\alpha, \quad z \in \mathbb{R}^{\mathcal{A}}_+, \quad n^{(z)}_\alpha(0) = z_\alpha, \quad \alpha \in \mathcal{A}.$$

If we assume

$$(1.8)\qquad \bar n_A > 0, \quad \bar n_a > 0, \quad \text{and} \quad S_{Aa} < 0 < S_{aA},$$

then $\bar n_\alpha$ is the equilibrium size of a monomorphic α-population and the system (1.7) has a unique stable equilibrium $(0, \bar n_a)$ and two unstable steady states $(\bar n_A, 0)$ and $(0,0)$. Thanks to Theorem 2.1 p. 456 in [11] we can prove that if $N^K_A(0)$ and $N^K_a(0)$ are of order K and K is large, the rescaled process $(N^K_A/K, N^K_a/K)$ is very close to the solution of (1.7) during any finite time interval. The invasion fitness S_{aA} corresponds to the *per capita* initial growth rate of the mutant a when it appears in a monomorphic population of individuals A at their equilibrium size $\bar n_A K$. Hence the dynamics of the allele a is very dependent on the properties of the system (1.7) and it is proven in [3] that under Condition (1.8) one mutant a has a positive probability to fix in the population and replace a wild type A. More precisely, if we use the convention

$$(1.9)\qquad \mathbb{P}^{(K)}(.) := \mathbb{P}(.|N^K_A(0) = \lfloor \bar n_A K \rfloor, N^K_a(0) = 1),$$

Equation (39) in [3] states that

$$\lim_{K\to\infty} \mathbb{P}^{(K)}(\mathrm{Fix}^K) = \frac{S_{aA}}{f_a} =: s, \tag{1.10}$$

where s is called the rescaled invasion fitness, and the extinction time of the A-population and the event of fixation of the a-allele are rigorously defined as follows:

$$T^K_{\mathrm{ext}} := \inf\left\{t \geq 0 : N^K_A(t) = 0\right\} \quad \text{and} \quad \mathrm{Fix}^K := \left\{T^K_{\mathrm{ext}} < \infty, N^K_a(T^K_{\mathrm{ext}}) > 0\right\}. \tag{1.11}$$

From this point onward, we fix d in $\mathbb{N}$. We aim at quantifying the effect of the selective sweep on the neutral diversity. Our method consists in tracing back the neutral genealogies of d individuals sampled uniformly at the end of the sweep (time T^K_{ext}) until time 0. Two event types (see Definition 4.1) may affect the relationships of the sampled neutral alleles: coalescences correspond to the merging of the neutral genealogies of two individuals at one or two neutral loci, and recombinations redistribute the selected and neutral alleles of one individual into two groups carried by its two parents. We will represent the neutral genealogies by a partition Θ^K_d which belongs to the set $\mathcal{P}^*_d$ of marked partitions of $\{(i,k), i \in \{1,...,d\}, k \in \{1,2\}\}$ with (at most) one block distinguished by the mark $*$, which will correspond to the descendants of the original mutant a. In this notation $(i,1)$ and $(i,2)$ are the neutral alleles at loci $N1$ and $N2$ of the ith sampled individual. Let us define rigorously the random partition Θ^K_d:

Definition 1.1. *Sample d individuals uniformly and without replacement at the end of the sweep (time T^K_{ext}). Follow the genealogies of the first and second neutral alleles of the i-th sampled individual, $(i,1)$ and $(i,2)$ for $i \in \{1,...,d\}$. Then the partition $\Theta^K_d \in \mathcal{P}^*_d$ is defined as follows: each block of the partition Θ^K_d is composed of all those neutral alleles which originate from the same individual alive at the beginning of the sweep; the block containing the descendants of the mutant a (if such a block exists) is distinguished by the mark $*$.*

We will show in Theorems 1 and 2 that when K is large the partition Θ^K_d belongs with a probability close to one to a subset Δ_d of $\mathcal{P}^*_d$, which is defined as follows:

Definition 1.2. *Δ_d is the subset of $\mathcal{P}^*_d$ consisting of those partitions whose unmarked blocks (if there are any) are either singletons or pairs of the form $\{(i,1),(i,2)\}$ for one $i \in \{1,...,d\}$.*

Example 1. In the example represented in Figure 1, the marked partition $\pi^{(ex)}$ belongs to Δ_d:

$$\pi^{(ex)} = \Big\{\{(1,1),(1,2),(2,1),(5,2)\}^*, \{(2,2)\}, \{(3,1),(3,2)\}, \{(4,1)\}, \{(4,2)\}, \{(5,1)\}\Big\}.$$

For a partition $\pi \in \mathcal{P}^*_d$, we define for some possible ancestral relationships the number of individuals in the sample whose two neutral loci are related in that particular way:

Definition 1.3. *Let $d \in \mathbb{N}$ and $\pi \in \mathcal{P}^*_d$. Then we set:*

$$\begin{aligned}
|\pi|_1 &= \#\{1 \leq i \leq d \text{ such that } (i,1) \text{ and } (i,2) \text{ belong to the marked block }\}\\
|\pi|_2 &= \#\{1 \leq i \leq d \text{ such that } (i,1) \text{ belongs to the marked block and } \{(i,2)\} \text{ is an unmarked block}\}\\
|\pi|_3 &= \#\{1 \leq i \leq d \text{ such that } (i,2) \text{ belongs to the marked block and } \{(i,1)\} \text{ is an unmarked block}\}\\
|\pi|_4 &= \#\{1 \leq i \leq d \text{ such that } \{(i,1),(i,2)\} \text{ is an unmarked block}\}\\
|\pi|_5 &= \#\{1 \leq i \leq d \text{ such that } \{(i,1)\} \text{ and } \{(i,2)\} \text{ are two distinct unmarked blocks}\}
\end{aligned}$$

To express the limit distribution of the partition Θ^K_d we need to introduce:

$$q_1 := e^{-\frac{f_a r_1 \log K}{S_{aA}}}, \quad q_2 := e^{-\frac{f_a r_2 \log K}{S_{aA}}}, \quad \bar{q}_2 := e^{-\frac{f_a r_2 \log K}{|S_{Aa}|}} \quad \text{and} \quad q_3 := \frac{r_1(q_2^{f_A/f_a} - q_1 q_2)}{r_1 + r_2(1 - f_A/f_a)}, \tag{1.12}$$

where the invasion fitnesses have been defined in (1.5). We did not make any assumption on the sign of $f_a(r_1+r_2) - f_A r_2$, but q_3 can be written in the form $\delta(e^{-\mu} - e^{-\nu})/(\nu - \mu)$ for $(\delta,\mu,\nu) \in \mathbb{R}^3_+$ so that it is well defined and non-negative. It is easy to check that $q_3 \leq 1$. The forms of q_1, q_2 and $\bar{q}_2$ are intuitive (see comments of Proposition 1).The form of q_3 is more complex to explain and results from a combination of different possible genealogical scenarios during the first phase. We now define five non-negative numbers $(p_k, 1 \leq k \leq 5)$ which will quantify the law of Θ^K_d for large K in Theorem 1:

$$\begin{gathered}
p_1 := q_1 q_2[1 - (1-q_1)(1-\bar{q}_2)], \quad p_2 := q_1[(1 - q_1 q_2) - q_2\bar{q}_2(1-q_1)],\\
p_3 := q_1 q_2(1-\bar{q}_2)(1-q_1), \quad p_4 := \bar{q}_2 q_3 \quad \text{and} \quad p_5 := (1-q_1)(1 - q_1 q_2(1-\bar{q}_2)) - \bar{q}_2 q_3.
\end{gathered} \tag{1.13}$$

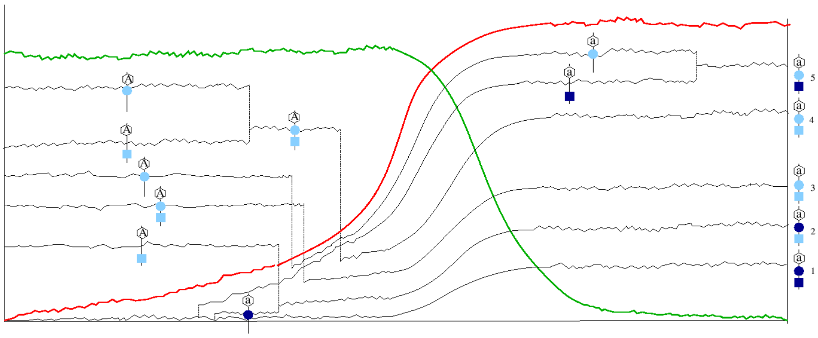

Figure 1. Example of genealogy for a 5-sample: dark blue neutral alleles originate from the mutant and light blue ones from an A-individual. We indicate the selected allele, A or a, associated with the neutral alleles during the sweep. It can change when a recombination occurs. Bold lines represent the A(green)- and a(red)-population sizes. In this example, the two neutral alleles of the first individual, the first neutral allele of the second individual and the second neutral allele of the fifth individual originate from the mutant; the two neutral alleles of the third individual originate from the same A-individual, whereas the two neutral alleles of the fourth individual originate from two distinct A-individuals.

Note that $\sum_{1\leq k\leq 5} p_k = 1$. Finally, we introduce an assumption which summarizes all the assumptions made in this work:

Assumption 1. $(N_A^K(0), N_a^K(0)) = (\lfloor \bar{n}_A K\rfloor, 1)$ *and Conditions* (1.1) *on the recombination probability and* (1.8) *on the equilibrium densities and fitnesses hold.*

With Definitions 1.1, 1.2 and 1.3 in mind, we can now state our main results:

Theorem 1 (Geometry $SL-N1-N2$). *Under Assumption 1, we have for every $\pi\in\mathcal{P}_d^*$*

$$\lim_{K\to\infty}\left|\mathbb{P}^{(K)}(\Theta_d^K=\pi|\text{Fix}^K) - \mathbf{1}_{\{\pi\in\Delta_d\}}{p_1}^{|\pi|_1}{p_2}^{|\pi|_2}{p_3}^{|\pi|_3}{p_4}^{|\pi|_4}{p_5}^{|\pi|_5}\right| = 0.$$

Notice that when K is large, Θ_d^K belongs to Δ_d with a probability close to one, and that

$$({p_1}^{|\pi|_1}{p_2}^{|\pi|_2}{p_3}^{|\pi|_3}{p_4}^{|\pi|_4}{p_5}^{|\pi|_5}, \pi\in\Delta_d)$$

is a probability on Δ_d (depending on K). Moreover, this result implies that the d sampled individuals have asymptotically independent neutral genealogies. With high probability, the neutral alleles of a given sampled individual i either originate from the first mutant a and belong to the marked block, or escape the sweep and originate from an A individual. In this case they belong to an unmarked block which is of the form $\{(i,1)\}$, $\{(i,2)\}$ or $\{(i,1),(i,2)\}$, according to Definition 1.3. As a consequence, if some neutral alleles of two distinct sampled individuals escape the sweep, they originate from distinct A-individuals with high probability. However, the genealogies of the two neutral alleles of a given individual are not independent. For example the probability that $(i,1)$ and $(i,2)$ escape the sweep is p_4+p_5; the probability that $(i,1)$ (resp. $(i,2)$) escapes the sweep is $p_3+p_4+p_5$ (resp. $p_2+p_4+p_5$), and for every $K\in\mathbb{N}$ such that $r_1\neq 0$

$$(p_3+p_4+p_5)(p_2+p_4+p_5) = (1-q_1)(1-q_1q_2) < (1-q_1)(1-q_1q_2+q_1q_2\bar{q}_2) = p_4+p_5.$$

This is due to the fact that if (backwards in time) a recombination first occurs between SL and $N1$, the neutral allele at $N2$, linked to $N1$, also escapes the sweep. As the term $q_1q_2\bar{q}_2$ does not tend to 0 when K goes to infinity under Condition (1.1), the only possibility to have an equality in the limit is the case where $r_1\log K \ll 1$ or in other words when the probability to see a recombination between SL and $N1$ is negligible.

Let us now consider the separated geometry, $N1-SL-N2$:

Theorem 2 (Geometry $N1 - SL - N2$). *Under Assumption 1, we have for every $\pi \in \mathcal{P}_d^*$*

$$\lim_{K\to\infty} \left| \mathbb{P}^{(K)}(\Theta_d^K = \pi | \mathrm{Fix}^K) - \mathbf{1}_{\{\pi \in \Delta_d\}} [q_1 q_2]^{|\pi|_1} [q_1(1-q_2)]^{|\pi|_2} [(1-q_1)q_2]^{|\pi|_3} [(1-q_1)(1-q_2)]^{|\pi|_5} \right| = 0.$$

Again the neutral genealogies of the d sampled individuals are asymptotically independent. Furthermore, we have independence between the neutral loci. Indeed Theorem 2 means that a neutral allele at locus Nk escapes the sweep with probability $1 - q_k$ independently of all other neutral alleles, including the allele at the other neutral locus of the same individual. This is due to the fact that in the separated geometry a recombination between SL and one neutral locus has no impact on the genetic background of the allele at the other neutral locus. Note in particular that there is no block of the form $\{(i,1),(i,2)\}$ in the limit partition, as the two neutral alleles have a very small probability to recombine at the same time.

2. Comparison with previous work

In [18] the authors gave an approximate sampling formula for the genealogy of one neutral locus during a selective sweep. The population evolved as a two-locus modified Moran model with recombination, selection, and in particular constant population size. They introduced the fitness s^{SD} of the mutant a as follows: when one of the iid exponential clocks of the living individuals rings, one picks two individuals uniformly at random (with replacement), one dies, and the other one gives birth. A replacement of an a-individual by an A-individual is rejected with probability s^{SD}. In this case, nothing happens. In [19], the author studied the one neutral locus version of the here presented model. It was shown that the ancestral relationships in a sample taken at the end of the sweep correspond to the ones derived in [18] when we equal the fitness of [18] and the rescaled invasion fitness $s^{SD} = S_{aA}/f_a$ and when we have the equality $|S_{Aa}|/f_A = S_{aA}/f_a$ (in this case the first and third phases have the same duration, $S_{aA} \log K/f_a$).

In [2], the author generalized the model introduced in [18] towards two neutral loci and used similar methods to derive a corresponding statement for the genealogy of a sample taken at the end of the sweep. If we however make the analogous comparison and try to match our result for the adjacent geometry with the statement from [2], we observe an interesting phenomenon: the probabilities of the different types of ancestry only coincide if the birth rates of a- and A-individuals are the same, that is, if $f_a = f_A$. In biology, the fitness describes the ability to both survive and reproduce, and can be defined by the average contribution of an individual with a given genotype to the gene pool of the next generation. Hence a mutation which affects the fitness of an individual in a given environment can either act on the fertility (f_α in our model), or on the death rate, intrinsic (D_α) or by competition ($C_{\alpha,\alpha'}$), or on both. Our result is comparable to that of [18] if the mutation only affects the death rate (and still if $s^{SD} = S_{aA}/f_a = |S_{Aa}|/f_A$).

In [17], instead of a birth and death process, the authors modeled the population with a structured coalescent. It is shown that this process can be approximated by a marked Yule tree where the different marks are realized by Poisson processes and indicate a recombination of one or two loci into the wild type background. The impact of the third phase is taken into account by a certain refinement prior to the beginning of the coalescent which leads to the same effect of splitting of the two neutral loci as it is seen here. We again find similarities with our results when $f_A = f_a$. In contrast, the techniques and precision used in [17] yield that coalescent events with A-individuals cannot be ignored, that is, there are neutral loci of different individuals from the sample which have the same type-A-ancestor. The structure of the sample is therefore different from our results here. Notice that it is also the case of the second approximate sampling formula stated in [18], which is more precise than the first one.

3. Dynamics of the sweep and couplings

3.1. **Description of the three phases.** We only need to focus on the trajectories of the population process where the mutant allele a goes to fixation and replaces the resident allele A. Champagnat has described these trajectories in [3] and in particular divided the sweep into three phases with distinct A- and a-population dynamics (see Figure 2). In the sequel, ε will be a positive real number independent of K, as small as needed for the different approximations to hold. Moreover, from this point onward we will write N_α (resp. $N_{\alpha\beta\gamma}$) instead of N_α^K (resp. $N_{\alpha\beta\gamma}^K$) and $\mathbb{P}$ instead of $\mathbb{P}^{(K)}$ for the sake of readability.

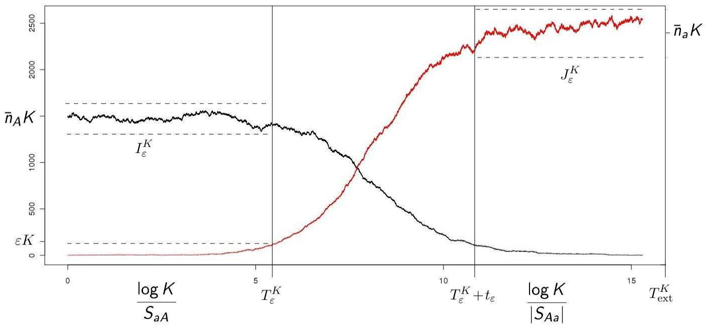

FIGURE 2. The three phases of a selective sweep The y-axis corresponds to population sizes (A in black, a in red), and the x-axis to the time. In this simulation, $K = 1000$, $(f_A, f_a) = (2, 3), D_\alpha = 0.5, \alpha \in \mathcal{A}$, $C_{\alpha,\alpha'} = 1, (\alpha, \alpha') \in \mathcal{A}^2$. We have also indicated some of the notations introduced in Section 3.1

First phase. The resident population size stays close to its equilibrium value $\bar{n}_A K$ as long as the mutant population size has not hit $\lfloor \varepsilon K \rfloor$: if we introduce the finite subset of $\mathbb{N}$

$$I_\varepsilon^K := \Big[K\Big(\bar{n}_A - 2\varepsilon\frac{C_{A,a}}{C_{A,A}}\Big), K\Big(\bar{n}_A + 2\varepsilon\frac{C_{A,a}}{C_{A,A}}\Big)\Big] \cap \mathbb{N}, \tag{3.1}$$

and the stopping times T_ε^K and S_ε^K, which denote respectively the hitting time of $\lfloor \varepsilon K \rfloor$ by the mutant population size and the exit time of I_ε^K by the resident population size,

$$T_\varepsilon^K := \inf\{t \geq 0, N_a(t) = \lfloor \varepsilon K \rfloor\} \quad \text{and} \quad S_\varepsilon^K := \inf\{t \geq 0, N_A(t) \notin I_\varepsilon^K\}, \tag{3.2}$$

then we can deduce from [3] (see Equations (A.5) and (A.6) in [19] for the details of the derivation) that the events Fix^K, $\{T_\varepsilon^K \leq S_\varepsilon^K\}$ and $\{T_\varepsilon^K < \infty\}$ are very close:

$$\limsup_{K\to\infty} \mathbb{P}^{(K)}(\{T_\varepsilon^K \leq S_\varepsilon^K\} \triangle \text{Fix}^K) \leq c\varepsilon, \quad \text{and} \quad \limsup_{K\to\infty} \mathbb{P}^{(K)}(\{T_\varepsilon^K < \infty\} \triangle \text{Fix}^K) \leq c\varepsilon, \tag{3.3}$$

for a finite c and ε small enough, where we recall convention (1.9). In this context, $\triangle$ is the symmetric difference: for two sets B and C, $B \triangle C = (B \cap C^c) \cup (C \cap B^c)$. From this point onwards, "first phase" will denote the time interval $[0, T_\varepsilon^K]$ when the a-population size is smaller than $\lfloor \varepsilon K \rfloor$.

Second phase. When N_A and N_a are of order K, the rescaled population process $(N_A/K, N_a/K)$ is well approximated by the Lotka-Volterra system (1.7). Moreover, under Condition (1.8) the system (1.7) has a unique attracting equilibrium $(0, \bar{n}_a)$ for initial conditions z satisfying $z_a > 0$, where $\bar{n}_a$ has been defined in (1.6). In particular, if we introduce for $(n_A, n_a) \in \mathbb{N}^2$ the notation,

$$\mathbb{P}_{(n_A,n_a)}(.) := \mathbb{P}(.|N_A(0) = n_A, N_a(0) = n_a), \tag{3.4}$$

then Theorem 3 (b) in [3] implies:

$$\lim_{K\to\infty} \sup_{z\in\Gamma} \mathbb{P}_{(\lfloor z_A K\rfloor, \lfloor z_a K\rfloor)}\Big(\sup_{0\leq t\leq t_\varepsilon, \alpha\in\mathcal{A}} \Big|\frac{N_\alpha(t)}{K} - n_\alpha^{(z)}(t)\Big| \geq \delta\Big) = 0, \tag{3.5}$$

for every $\delta > 0$, where

$$\Gamma := \Big\{z \in \mathbb{R}_+^{\mathcal{A}}, \lfloor z_A K\rfloor \in I_\varepsilon^K, z_a \in [\varepsilon/2, \varepsilon]\Big\}, \tag{3.6}$$

$$\begin{aligned} t_\varepsilon(z) &:= \inf\big\{s \geq 0, \forall t \geq s, n_A^{(z)}(t) \in [0, \varepsilon^2/2], n_a^{(z)}(t) \in [\bar{n}_a - \varepsilon/2, \bar{n}_a + \varepsilon/2]\big\}, \\ t_\varepsilon &:= \sup\{t_\varepsilon(z), z \in \Gamma\} < \infty. \end{aligned} \tag{3.7}$$

In the sequel, "second phase" will denote the time interval $[T_\varepsilon^K, T_\varepsilon^K + t_\varepsilon]$ when the population process is close to the solution of the system (1.7).

Third phase. Equation (3.5) also implies that

$$\lim_{K\to\infty} \mathbb{P}\Big(\tfrac{N_A(T_\varepsilon^K+t_\varepsilon)}{K} \in [\omega_1,\omega_2], \Big|\tfrac{N_a(T_\varepsilon^K+t_\varepsilon)}{K} - \bar{n}_a\Big| \leq \varepsilon, \ \Big|\Big(\tfrac{N_A(T_\varepsilon^K)}{K}, \tfrac{N_a(T_\varepsilon^K)}{K}\Big) \in \Gamma\Big) = 1, \tag{3.8}$$

where

$$2\omega_1 := \inf \{n_A^{(z)}(t_\varepsilon), z\in\Gamma\} > 0, \quad \text{and} \quad \omega_2 := 2\sup\{n_A^{(z)}(t_\varepsilon), z \in \Gamma\} \leq \varepsilon^2. \tag{3.9}$$

The "third phase", which corresponds to the time interval $[T_\varepsilon^K + t_\varepsilon, T_{\text{ext}}^K]$, can be seen as the symmetric counterpart of the first phase, where the roles of A and a are interchanged: during the extinction of the A-population, the a-population size stays close to its equilibrium value $\bar{n}_a K$.

Let us introduce the positive real number $M'' := 3 + (f_a + C_{a,A})/C_{a,a}$ and the finite subset of $\mathbb{N}$

$$J_\varepsilon^K := \Big[K\Big(\bar{n}_a - M''\varepsilon\Big), K\Big(\bar{n}_a + M''\varepsilon\Big)\Big] \cap \mathbb{N}. \tag{3.10}$$

The times $T_u^{(K,A)}$ and $S_\varepsilon^{(K,a)}$ are two stopping times for the process restarted after the second phase and denote respectively the hitting times of $\lfloor uK \rfloor$ by the A-population for $u \in \mathbb{R}_+$, and the exit time of J_ε^K by the a-population during the third phase,

$$T_u^{(K,A)} := \inf\{t \geq 0, N_A(T_\varepsilon^K + t_\varepsilon + t) = \lfloor uK \rfloor\}, \quad S_\varepsilon^{(K,a)} := \inf\{t\geq 0, N_a(T_\varepsilon^K + t_\varepsilon + t) \notin J_\varepsilon^K\}. \tag{3.11}$$

If we define the event

$$\mathcal{N}_\varepsilon^K := \{T_\varepsilon^K \leq S_\varepsilon^K\} \cap \Big\{\tfrac{N_A(T_\varepsilon^K+t_\varepsilon)}{K} \in [\omega_1,\omega_2], \Big|\tfrac{N_a(T_\varepsilon^K+t_\varepsilon)}{K} - \bar{n}_a\Big| \leq \varepsilon\Big\}, \tag{3.12}$$

we get from the proof of Lemma 3 in [3] that for a finite c and ε small enough,

$$\limsup_{K\to\infty} \Big\{\mathbb{P}(\text{Fix}^K \vartriangle [\mathcal{N}_\varepsilon^K \cap \{T_0^{(K,A)} < T_\varepsilon^{(K,A)} \wedge S_\varepsilon^{(K,a)}\}]) + \mathbb{P}(\text{Fix}^K \vartriangle [\mathcal{N}_\varepsilon^K \cap \{T_0^{(K,A)} < T_\varepsilon^{(K,A)}\}])\Big\} \leq c\varepsilon. \tag{3.13}$$

To summarize, the fixation event Fix^K is very close to the following succession of events:

- The a-population size hits $\lfloor \varepsilon K \rfloor$ before the A-population size has escaped the vicinity of its equilibrium I_ε^K (first phase)
- The rescaled population process N/K is close to the deterministic competitive Lotka-Volterra system during the second phase
- The A-population size gets extinct before hitting $\lfloor \varepsilon K \rfloor$ and before the a-population size has escaped the vicinity of its equilibrium J_ε^K (third phase)

3.2. **Couplings for the first and third phases.** We are interested in the law of the neutral genealogies on the event Fix^K. Equations (3.3) and (3.13) imply that it is enough to concentrate our attention on the event $\mathcal{N}_\varepsilon^K \cap \{T_0^{(K,A)} < T_\varepsilon^{(K,A)}\}$, but the dynamics of the population process N conditionally on this event is complex to study. Indeed it boils down to studying the dynamics of a process conditioned on a future event ($\{T_\varepsilon^K \leq S_\varepsilon^K\}$ for the first phase and $\{T_0^{(K,A)} < T_\varepsilon^{(K,A)}\}$ for the third one). Hence the idea is to couple the population process with two processes, $\tilde{N}$ and $\tilde{\tilde{N}}$, whose laws are easier to study. These processes will satisfy:

$$\limsup_{K\to\infty} \mathbb{P}(\{\exists t \leq T_\varepsilon^K, N(t) \neq \tilde{N}(t)\}, T_\varepsilon^K < \infty) \leq c\varepsilon. \tag{3.14}$$

and

$$\limsup_{K\to\infty} \mathbb{P}(\{\exists\ 0 \leq t - (T_\varepsilon^K + t_\varepsilon) \leq T_0^{(K,A)}, N(t) \neq \tilde{\tilde{N}}(t)\}, T_0^{(K,A)} < T_\varepsilon^{(K,A)} | \mathcal{N}_\varepsilon^K < \infty) \leq c\varepsilon. \tag{3.15}$$

Let α be in $\mathcal{A}$ and n be in $\mathbb{N}^{\mathcal{E}}$. Denote $n^{(\alpha)}$ the α component of the population state:

$$n^{(\alpha)} = \sum_{(\beta,\gamma)\in\mathcal{B}\times\mathcal{C}} n_{\alpha\beta\gamma} e_{\alpha\beta\gamma}, \tag{3.16}$$

where $(e_{\alpha\beta\gamma}, (\alpha,\beta,\gamma)\in\mathcal{E})$ is the canonical basis of $\mathbb{R}^{\mathcal{E}}$. We are now able to introduce a process needed to describe the couplings:

Definition 3.1. *We denote by* Moran process of type α with recombination r_2 *a process* $MR_\alpha^{(n^{(\alpha)})}$ *with values in* $\mathbb{N}^{\alpha\times\mathcal{B}\times\mathcal{C}}$, *initial state* $n^{(\alpha)}$, *and the following dynamics:*

- *After an exponential time with parameter $f_\alpha \bar{n}_\alpha K$ we pick uniformly and with replacement three individuals and draw a Bernoulli variable R with parameter r_2*
- *The first individual dies, the second one gives birth to an individual carrying its alleles at loci SL and $N1$, the third one is the potential second parent*
- *If $R = 0$, there is no recombination and the allele at locus $N2$ of the newborn is also inherited from the second individual; if $R = 1$ there is a recombination between $N1$ and $N2$ and the newborn inherits its second neutral allele from the third individual*
- *We again draw an exponential variable with parameter $f_\alpha \bar{n}_\alpha K$ and restart the procedure*

Coupling with $\tilde{N}$: N and $\tilde{N}$ are equal up to time S_ε^K; after this time the A individuals in the population process $\tilde{N}$ follow a Moran process with recombination independent of the a-individuals. Let $\underline{\mathfrak{bc}}$ be in $\mathcal{B} \times \mathcal{C}$. We let the a-population evolve as if the a-individuals were interacting with $\tilde{N}_A(s)$ individuals with genotype $A\underline{\mathfrak{bc}}$:

$$\tilde{N}(t) = \mathbf{1}_{t<S_\varepsilon^K} N(t) + \mathbf{1}_{t \geq S_\varepsilon^K} \Big(MR_A^{(N^{(A)}(S_\varepsilon^K))}(t - S_\varepsilon^K) + \sum_{(\beta,\gamma)\in\mathcal{B}\times\mathcal{C}} e_{a\beta\gamma} \tag{3.17}$$

$$\int_{S_\varepsilon^K}^t \int_{\mathbb{R}_+} \Big[Q_{\beta\gamma}^{(1)}(ds, d\theta) \mathbf{1}_{\{0<\theta\leq b_{a\beta\gamma}^K(\tilde{N}_A(s-)e_{A\underline{\mathfrak{bc}}}, \tilde{N}^{(a)}(s-))\}} - Q_{\beta\gamma}^{(2)}(ds, d\theta) \mathbf{1}_{\{0<\theta\leq d_{a\beta\gamma}^K(\tilde{N}_A(s-)e_{A\underline{\mathfrak{bc}}}, \tilde{N}^{(a)}(s-))\}} \Big] \Big),$$

where $MR_A^{(N^{(A)})}$ has been defined in Definition 3.1 and $(Q_{\beta\gamma}^{(i)}, i \in \{1,2\}, (\beta,\gamma) \in \mathcal{B} \times \mathcal{C})$ are independent Poisson Point processes with density $dsd\theta$, also independent of $MR^{(N^{(A)})}$. The reason for the construction of such a coupling is that we need to control the A-population size and the number of births of A-individuals during the first phase in Section 5. With the process $\tilde{N}$ such control is achieved easier.

Coupling with $\tilde{\tilde{N}}$: we assume that $\mathcal{N}_\varepsilon^K$ from (3.12) holds; N and $\tilde{\tilde{N}}$ are equal up to time $T_\varepsilon^K + t_\varepsilon + S_\varepsilon^{(K,a)} \wedge T_\varepsilon^{(K,A)}$. Then the a-individuals in the population process $\tilde{\tilde{N}}$ follow a Moran process with recombination independent of the A-individuals, and each $A\beta\gamma$-population evolves as a birth and death process with individual birth and death rates f_A and $f_A + |S_{Aa}|$, independent of the a-individuals and the $A\beta'\gamma'$-populations with $(\beta,\gamma) \neq (\beta',\gamma')$:

$$\tilde{\tilde{N}}(T_\varepsilon^K + t_\varepsilon + t) = \mathbf{1}_{t<S_\varepsilon^{(K,a)}} N(T_\varepsilon^K + t_\varepsilon + t) + \mathbf{1}_{t\geq S_\varepsilon^{(K,a)}} \Big(MR_a^{(N^{(a)}(S_\varepsilon^{(K,a)}))}(t - S_\varepsilon^{(K,a)}) + \tag{3.18}$$

$$+ \sum_{(\beta,\gamma)\in\mathcal{B}\times\mathcal{C}} e_{A\beta\gamma} \Big[\int_{T_\varepsilon^K + t_\varepsilon + S_\varepsilon^{(K,a)}}^{T_\varepsilon^K + t_\varepsilon + t} \int_{\mathbb{R}_+} Q_{\beta\gamma}(ds, d\theta) \Big\{ \mathbf{1}_{\{0<\theta\leq f_A \tilde{\tilde{N}}_{A\beta\gamma}(s-)\}} - \mathbf{1}_{\{0<\theta - f_A \tilde{\tilde{N}}_{A\beta\gamma}(s-) \leq (f_A + |S_{Aa}|)\tilde{\tilde{N}}_{A\beta\gamma}(s-)\}} \Big\} \Big] \Big),$$

where $MR_a^{(N^{(a)})}$ has been defined in Definition 3.1 and is independent of the sequence of independent Poisson measures $(Q_{\beta\gamma}, (\beta,\gamma) \in \mathcal{B} \times \mathcal{C})$, with intensity $dsd\theta$. The a-population size and the number of births of a-individuals will be easy to control for the process $\tilde{\tilde{N}}$ during the third phase, and again we will need such control in Section 5.

Inequality (3.14) follows from (3.3). Moreover, from the proof of Lemma 3 in [3] we know that

$$\liminf_{K\to\infty} \mathbb{P}(T_0^{(K,A)} < T_\varepsilon^{(K,A)} \wedge S_\varepsilon^{(K,a)} | \mathcal{N}_\varepsilon^K) \geq 1 - c\varepsilon$$

for a finite c and ε small enough. Adding (3.13) we get that (3.15) is also satisfied. Hence we will study the processes $\tilde{N}$ and $\tilde{\tilde{N}}$ and deduce properties of the dynamics of the process N during the first and third phases.

4. Proofs of the main results

As the proof of Theorem 2 is simpler than this of Theorem 1, and follows essentially the same ideas, we only prove Theorem 1.

4.1. **Events impacting the genealogies in each phase.** Let us now summarize the results on the genealogies for the three successive phases of the sweep that we will derive in Sections 6 and 7.

First phase: As explained in the previous section, we work with the process $\tilde{N}$ to study the first phase. Let us introduce the jump times of $\tilde{N}$:

$$\tau_0^K = 0 \quad \text{and} \quad \tau_m^K = \inf\{t > \tau_{m-1}^K, \tilde{N}(t) \neq \tilde{N}(\tau_{m-1}^K)\}, \quad m \geq 1. \tag{4.1}$$

The number of jumps during the first phase is denoted by $J^K(1)$:

$$J^K(1) := \inf\{m \in \mathbb{N}, \tilde{N}_a(\tau_m^K) = \lfloor \varepsilon K \rfloor\}. \tag{4.2}$$

Coalescence and recombination events are defined as follows (see Figure 3):

Definition 4.1. *Sample two distinct individuals at time τ_m^K and denote $\alpha\beta\gamma$ and $\alpha'\beta'\gamma'$ their type.*

We say that β and β' coalesce at time τ_m^K if they are carried by two distinct individuals at time τ_m^K and by the same individual at time τ_{m-1}^K. Seen forwards in time it corresponds to a birth and hence a copy of the neutral allele. Seen backwards in time it corresponds to the fusion of two neutral alleles into one, carried by one parent of the newborn. We define in the same way coalescent events at locus $N2$ (resp. loci $N1$ and $N2$) for alleles γ and γ' (resp. allele pairings (β,γ) and (β',γ')).

We say that β (and/or γ) recombines at time τ_m^K from the α- to the α'-population if the individual carrying the allele β (and/or γ) at time τ_m^K is a newborn, carries the allele α inherited from it first parent, and has inherited its allele β (and/or γ) from a different individual carrying allele α'.

We are only interested in recombinations which entail new associations of alleles. In particular we will not consider the simultaneous recombinations of a pair (β,γ) within the α-population.

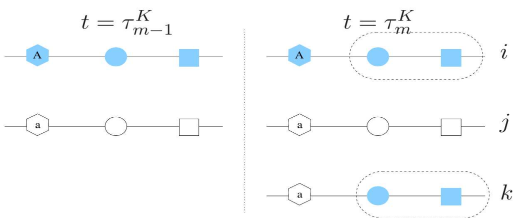

FIGURE 3. Illustration of Definition 4.1: the newborn (individual k) has inherited the selected allele from its "white" parent and the two neutral alleles from its "blue" parent; hence the encircled neutral loci (of individuals i and k) coalesce at time τ_m^K. In terms of recombinations, the two neutral loci of the newborn individual recombine at time τ_m^K from the a- to the A-population

Let us now describe the genealogical scenarios which modify the ancestral relationships between the neutral alleles of one individual and occur with positive probability when K is large. We first focus on the first phase and pick uniformly an individual i from the a-population at time $\tilde{T}_\varepsilon^K$. We introduce:

$NR(i)^{(1)}$: there is no recombination into the A-population affecting $(i,1)$ or $(i,2)$ and both neutral loci of the i-individual originate from the first mutant,

$R2(i)^{(1)}$: only the neutral allele $(i,2)$ is affected by a recombination with the A-population, hence $(i,1)$ originates from the first mutant and $(i,2)$ from an A-individual,

$R12(i)^{(1)}$: one recombination between SL and $N1$ from the a- into the A-population occurs and both neutral alleles $(i,1)$ and $(i,2)$ originate from the same A-individual,

$[2,1]_{A,i}^{rec}$: first (backwards in time) $(i,2)$ recombines into the A-population, then $(i,1)$ recombines into the A-population and connects to a different individual than $(i,2)$.

$[12,2]_{A,i}^{rec}$: first (backwards in time) the tuple $\{(i,1),(i,2)\}$ recombines into the A-population, then a second recombination splits the two neutral loci inside the A-population.

$R1|2(i)^{(1,ga)}$: $[2,1]_{A,i}^{rec} \cup [12,2]_{A,i}^{rec}$ (see Figure 4)

Finally, we introduce a conditional probability for the process $\tilde{N}$:

$$\mathbb{P}^{(1)}(.) = \mathbb{P}(.|J^K(1) < \infty), \tag{4.3}$$

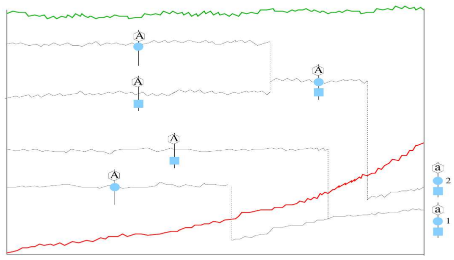

FIGURE 4. Illustration of events $[2,1]^{rec}_{A,i}$ (individual 1) and $[12,2]^{rec}_{A,i}$ (individual 2)

where $J^K(1)$ has been defined in (4.2). Hence, recalling the definition of (q_1, q_2, q_3) in (1.12) we will prove in Section 6:

Proposition 1 (Neutral genealogies during the first phase)**.** *Let i be an a-individual sampled uniformly at the end of the first phase (time $\tilde{T}^K_\varepsilon$). Under Assumption 1, there exist two finite constants c and ε_0 such that for every $\varepsilon \leq \varepsilon_0$,*

$$\limsup_{K\to\infty} \Big\{ \Big|\mathbb{P}^{(1)}(NR(i)^{(1)}) - q_1 q_2\Big| + \Big|\mathbb{P}^{(1)}(R2(i)^{(1)}) - q_1(1-q_2)\Big| + \Big|\mathbb{P}^{(1)}(R12(i)^{(1)}) - q_3\Big| + \Big|\mathbb{P}^{(1)}(R1|2(i)^{(1,ga)}) - (1-q_1-q_3)\Big|\Big\} \leq c\varepsilon.$$

For large K, the sum of the four probabilities of Proposition 1 equals one up to a constant times ε. Hence, in the limit we only observe the events described on page 10. The probabilities of the first two events are quite intuitive: broadly speaking, the probability to have no recombination at a birth event is $1-r_1-r_2$, the birth rate is f_a and the duration of the first phase is $\log K/S_{aA}$. Hence under $\mathbb{P}^{(1)}$, the probability of the event $NR(i)^{(1)}$ is approximately

$$(1-(r_1+r_2))^{f_a \log K/S_{aA}} \sim \exp(-(r_1+r_2))^{f_a \log K/S_{aA}} = q_1 q_2.$$

Similarly the probability to have no recombination between SL and $N1$ is close to q_1 and subtracting the probability of $NR(i)^{(1)}$ we get this of $R2(i)^{(1)}$. The probabilities of $R12(i)^{(1)}$ and $R1|2(i)^{(1,ga)}$ are more involved. The proofs rely on a fine study of the different possible scenarios.

Second phase: We work with the process N to study the second phase. The latter one has a duration of order 1, and the recombination probabilities are negligible with respect to one (Condition (1.1)). Consequently, no event impacting the genealogies of the neutral loci occurs during the second phase. More precisely, let us sample uniformly two distinct a-individuals i and j at the end of the second phase (time $T^K_\varepsilon + t_\varepsilon$) and introduce the events:

$$\begin{aligned} NR(i)^{(2)}: &\quad \text{there is no recombination affecting } (i,1) \text{ or } (i,2),\\ NC(i,j)^{(2)}: &\quad \text{there is no coalescence between the neutral genealogies of } i \text{ and } j. \end{aligned}$$

Then we have the following result, which will be proven in Section 7.

Proposition 2 (Neutral genealogies during the second phase)**.** *Let i and j be two distinct a-individuals sampled uniformly at the end of the second phase (time $T^K_\varepsilon + t_\varepsilon$). Then under Assumption 1,*

$$\lim_{K\to\infty} \mathbb{P}(NR(i)^{(2)} \cap NC(i,j)^{(2)} | T^K_\varepsilon \leq S^K_\varepsilon) = 1.$$

Third phase: Finally, we focus on the process $\tilde{\tilde{N}}$. When K is large, there is only one event occurring with positive probability during the third phase which may modify the ancestry of the neutral alleles of an individual i sampled at the end of the sweep in the adjacent geometry:

$$R2(i)^{(3,ga)}: \quad \text{a recombination between loci } N1 \text{ and } N2 \text{ occurs and separates } (i,1) \text{ and } (i,2) \text{ within the } a\text{-population,} \tag{4.4}$$

Indeed, if we also define the events

$NR(i)^{(3)}$: there is no recombination affecting $(i,1)$ or $(i,2)$ and they both originate from the same a-individual at the end of the second phase

$NC(i,j)^{(3)}$: defined as $NC(i,j)^{(2)}$ for two distinct individuals sampled uniformly at the end of the sweep,

and the conditional probability for the process $\tilde{\tilde{N}}$::

$$\mathbb{P}^{(3)}(.) := \mathbb{P}(.|\mathcal{N}_\varepsilon^K, \tilde{\tilde{T}}_0^{(K,A)} < \tilde{\tilde{T}}_\varepsilon^{(K,A)}), \tag{4.5}$$

where $\tilde{\tilde{T}}_0^{(K,A)}$ and $\tilde{\tilde{T}}_\varepsilon^{(K,A)}$ are the analogs of $T_0^{(K,A)}$ and $T_\varepsilon^{(K,A)}$ (defined in (3.11)) for the process $\tilde{\tilde{N}}$, then we will prove in Section 7:

Proposition 3 (Neutral genealogies during the third phase)**.** *Let i and j be two distinct a-individuals sampled uniformly at the end of the sweep. Under Assumption 1, there exist two finite constants c and ε_0 such that for every $\varepsilon \le \varepsilon_0$,*

$$\limsup_{K\to\infty} \Big\{ \Big|\mathbb{P}^{(3)}(R2(i)^{(3,ga)}) - (1-\bar{q}_2)\Big| + \Big|\mathbb{P}^{(3)}(NR(i)^{(3)}) - \bar{q}_2\Big| + \Big|\mathbb{P}^{(3)}(NC(i,j)^{(3)}) - 1\Big|\Big\} \le c\sqrt{\varepsilon}.$$

In particular, there is no recombination with the A-population during the third phase. As for the probabilities of the first two events in the Proposition 1 this result is quite intuitive, as the duration of the third phase is close to $\log K/|S_{Aa}|$.

Independence: Finally we again consider the population process N and state a proposition which enables us to give the statement of Theorem 1 independently for all sampled individuals, that is, jointly for the whole sample. To this aim, let us introduce a partition $\Theta_d^{(K,1)} \in \mathcal{P}_d^*$ which is the analog of Θ_d^K where the d individuals are sampled at the end of the first phase and not at the end of the sweep. Recall Definitions 1.2 and 1.3, and denote by $|R2^{(3,ga)}|_d$ (resp. $|NR^{(3)}|_d$) the number of a-individuals in a d-sample taken at the end of the sweep whose neutral alleles originate from two distinct a-individuals (resp. from the same a-individual) at the beginning of the third phase. Then we have the following result:

Proposition 4. *Let Assumption 1 hold. Then there exist two finite constants c and ε_0 such that for every $\varepsilon \le \varepsilon_0$, the ancestral relationships of a d-sample taken at the end of the first phase (time T_ε^K) satisfy for every $(m_k, 1 \le k \le 4) \in \mathbb{Z}_+^4$:*

$$\limsup_{K\to\infty} \Big| \mathbb{P}(|\Theta_d^{(K,1)}|_k = m_k, 1 \le k \le 4 | T_\varepsilon^K \le S_\varepsilon^K) - \mathbf{1}_{\{m_1+m_2+m_3+m_4=d\}} \frac{d!}{m_1!m_2!m_3!m_4!}(q_1q_2)^{m_1}(q_1(1-q_2))^{m_2} q_3^{m_3}(1-q_1-q_3)^{m_4} \Big| \le c\varepsilon.$$

In the same way, the neutral genealogy of a d-sample taken at the end of the sweep satisfies for every $(m_k, 1 \le k \le 2) \in \mathbb{Z}_+^2$:

$$\limsup_{K\to\infty} \Big| \mathbb{P}((|R2^{(3,ga)}|_d, |NR^{(3)}|_d) = (m_1,m_2)|\mathcal{N}_\varepsilon^K) - \mathbf{1}_{\{m_1+m_2=d\}} \frac{d!}{m_1!m_2!}(1-\bar{q}_2)^{m_1}\bar{q}_2^{m_2} \Big| \le c\varepsilon.$$

The Proposition 4 is a key result: we only need to focus on individual neutral genealogies to get general results on the genealogy of a d-sample with respect to the neutral loci. It will be proven in Section 8.

4.2. **Proof of Theorem 1.** Let i be an individual sampled uniformly at the end of the sweep. The idea of the proof is the following: in a first step, we list certain compositions of coalescent and recombination events leading to specific ancestral relationships which could be described by blocks of a partition of Δ_d. Then we approximate the probabilities of the described events and prove that these probabilities sum to one up to a constant times $\sqrt{\varepsilon}$ for some fixed small ε. This shows that in the limit for large K the neutral genealogy of the individual i belongs to those described on page 10 with a probability close to one. In a second step we use Proposition 4 to treat the neutral genealogies of the d sampled individuals independently.

i) We consider two possible trajectories such that the alleles at both neutral loci originate from the mutant: either the two neutral loci separate inside the a-population during the third phase and coalesce during the first phase, or they stay in the a-population and do not separate during the whole sweep (see individual 1 in Figure 1):

$$\Big(R2(i)^{(3,ga)} \cap NR(i1)^{(2)} \cap NR(i2)^{(2)} \cap NC(i1,i2)^{(2)} \cap [NR(i1)^{(1)} \sqcup R2(i1)^{(1)}] \cap NR(i2)^{(1)}\Big) \bigsqcup \Big(NR(i)^{(3)} \cap NR(i)^{(2)} \cap NR(i)^{(1)}\Big),$$

where $\sqcup$ is the disjoint union and we denote by $i1$ and $i2$ the labels of the parents of the first and second neutral loci of i, respectively, at the end of the second phase (the way we label the a-individuals has no importance as they are exchangeable).

ii) We consider two possible trajectories such that $(i,1)$ originates from the mutant and $(i,2)$ originates from some A-individual

$$\Big(R2(i)^{(3,ga)} \cap NR(i1)^{(2)} \cap NR(i2)^{(2)} \cap NC(i1,i2)^{(2)} \cap [NR(i1)^{(1)} \cup R2(i1)^{(1)}] \cap [R12(i2)^{(1)} \sqcup R1|2(i2)^{(1)} \sqcup R2(i2)^{(1)}]\Big) \bigsqcup \Big(NR(i)^{(3)} \cap NR(i)^{(2)} \cap R2(i)^{(1)}\Big).$$

The first bracket considers a separation of the two neutral loci during the third phase. As a consequence, the fate of the first neutral locus of individual $i2$ during the first phase has no consequence on the neutral genealogy of i. This is why we consider the event $\{R12(i2)^{(1)} \sqcup R1|2(i2)^{(1)} \sqcup R2(i2)^{(1)}\}$ and not only $\{R2(i2)^{(1)}\}$. The second bracket corresponds to individual 2 in Figure 1.

iii) We consider one possible trajectory such that $(i,1)$ originates from some A-individual and $(i,2)$ originates from the mutant (see individual 5 in Figure 1)

$$R2(i)^{(3,ga)} \cap NR(i1)^{(2)} \cap NR(i2)^{(2)} \cap NC(i1,i2)^{(2)} \cap [R12(i1)^{(1)} \sqcup R1|2(i1)^{(1,ga)}] \cap NR(i2)^{(1)}$$

iv) We consider one possible trajectory such that $(i,1)$ and $(i,2)$ originate from the same A-individual (see individual 3 in Figure 1)

$$NR(i)^{(3)} \cap NR(i)^{(2)} \cap R12(i)^{(1)}$$

v) Finally, we consider two possible trajectories such that $(i,1)$ and $(i,2)$ originate from distinct A-individuals (see individual 4 in Figure 1 for the second bracket):

$$\Big(R2(i)^{(3,ga)} \cap NR(i1)^{(2)} \cap NR(i2)^{(2)} \cap NC(i1,i2)^{(2)} \cap [R12(i1)^{(1)} \sqcup R1|2(i1)^{(1)}] \cap [R12(i2)^{(1)} \sqcup R1|2(i2)^{(1)} \cup R2(i2)^{(1)}]\Big) \bigsqcup \Big(NR(i)^{(3)} \cap NR(i)^{(2)} \cap R1|2(i)^{(1,ga)}\Big).$$

Thanks to (3.3), and (3.13) to (3.15) we know that for all non negligible measurable events $C^{(1)}$, $C^{(2)}$ and $C^{(3)}$ occurring during the first, second and third phase respectively,

$$\mathbb{P}(C^{(1)}, C^{(2)}, C^{(3)}, \text{Fix}^K) = \mathbb{P}(C^{(1)}, C^{(2)}, C^{(3)}, \mathcal{N}^K_\varepsilon, \{T_0^{(K,A)} < T_\varepsilon^{(K,A)} \wedge S_\varepsilon^{(K,A)}\}) + O_K(\varepsilon) \tag{4.6}$$

where $O_K(\varepsilon)$ is a function of K and ε satisfying

$$\limsup_{K\to\infty} |O_K(\varepsilon)| \leq c\varepsilon, \tag{4.7}$$

for $\varepsilon \leq \varepsilon_0$ where ε_0 and c are finite. Using the same inequalities we can decompose the right hand side of (4.6) as follows

$$\mathbb{P}(C^{(1)}, \{T_\varepsilon^K < \infty\}) + \mathbb{P}(C^{(2)}, \{\tfrac{N_A(T_\varepsilon^K + t_\varepsilon)}{K} \in [\omega_1, \omega_2], |\tfrac{N_a(T_\varepsilon^K + t_\varepsilon)}{K} - \bar{n}_a| \leq \varepsilon\} | C^{(1)}, \{T_\varepsilon^K \leq S_\varepsilon^K\}) + \mathbb{P}(C^{(3)}, \{T_0^{(K,A)} < T_\varepsilon^{(K,A)} \wedge S_\varepsilon^{(K,A)}\} | C^{(1)}, C^{(2)}, \mathcal{N}_\varepsilon^K) + O_K(\varepsilon).$$

Then from (3.14) we get

$$\mathbb{P}(C^{(1)}, \{T_\varepsilon^K < \infty\}) = \mathbb{P}^{(1)}(\tilde{C}^{(1)}) \mathbb{P}(T_\varepsilon^K < \infty) + O_K(\varepsilon),$$

from (3.8)

$\mathbb{P}(C^{(2)},\{\frac{N_A(T_\varepsilon^K+t_\varepsilon)}{K}\in[\omega_1,\omega_2],|\frac{N_a(T_\varepsilon^K+t_\varepsilon)}{K}-\bar{n}_a|\leq\varepsilon\}|C^{(1)},\{T_\varepsilon^K\leq S_\varepsilon^K\})=\mathbb{P}(C^{(2)}|C^{(1)},\{T_\varepsilon^K\leq S_\varepsilon^K\})+O_K(\varepsilon)$,

and from (3.13) and (3.15)

$$\mathbb{P}(C^{(3)},\{T_0^{(K,A)}<T_\varepsilon^{(K,A)}\wedge S_\varepsilon^{(K,A)}\}|C^{(1)},C^{(2)},\mathcal{N}_\varepsilon^K)=\mathbb{P}^{(3)}(\tilde{\tilde{C}}^{(3)}|C^{(1)},C^{(2)})+O_K(\varepsilon),$$

where $\tilde{C}^{(1)}$ (resp. $\tilde{\tilde{C}}^{(3)}$) corresponds to the event $C^{(1)}$ (resp. $C^{(3)}$) expressed in terms of the process $\tilde{N}$ (resp. $\tilde{\tilde{N}}$). Putting everything together we finally obtain

$$\mathbb{P}(C^{(1)},C^{(2)},C^{(3)}|\mathrm{Fix}^K)=\mathbb{P}^{(3)}(\tilde{\tilde{C}}^{(3)}|C^{(1)},C^{(2)})\mathbb{P}(C^{(2)}|C^{(1)},\{T_\varepsilon^K\leq S_\varepsilon^K\})\mathbb{P}^{(1)}(\tilde{C}^{(1)})+O_K(\varepsilon). \tag{4.8}$$

By applying Propositions 1, 2, 3 and 4 we then can calculate the probabilities corresponding to the five sets from Definition 1.3 and get the probabilities $(p_k, 1\leq k\leq 5)$ defined in (1.13), which sum to one. Let us detail the calculations for the case i): by applying (4.8), Proposition 2 and the Markov property, the probability to see one of the two trajectories described in i) is

$$\begin{aligned}\mathcal{P}(i,1)=\mathbb{P}^{(3)}(R2(i)^{(3,ga)})\mathbb{P}^{(1)}([NR(i)^{(1)}\sqcup R2(i)^{(1)}]\cap NR(j)^{(1)})&\\+\mathbb{P}^{(3)}(NR(i)^{(3)})\mathbb{P}^{(1)}(NR(i)^{(1)})+O_K(\varepsilon),&\end{aligned} \tag{4.9}$$

where i and j are two distinct individuals (exchangeability). But thanks to Proposition 4 we know that the neutral genealogies of individuals i and j are nearly independent. Hence adding Proposition 1 leads to

$$\mathbb{P}^{(1)}([NR(i)^{(1)}\sqcup R2(i)^{(1)}]\cap NR(j)^{(1)})=(q_1q_2+q_1(1-q_2))q_1q_2+O_K(\varepsilon).$$

Applying Propositions 1 and 3 in (4.9) yields

$$\mathcal{P}(i,1)=(1-\bar{q}_2)q_1^2q_2+\bar{q}_2q_1q_2+O_K(\sqrt{\varepsilon})=p_1+O_K(\sqrt{\varepsilon}),$$

where we recall the definition of p_1 in (1.13).

Finally, we get the asymptotic independence of the neutral genealogies of the d sampled individuals during the first and third phases by applying the multinomial version of the de Finetti Representation Theorem (see [8] Chapter 4 for a simple proof) to the result of Proposition 4. The asymptotic independence during the second phase follows from Proposition 2 as, with high probability, nothing happens.

5. Number of births and deaths during the selective sweep

In this section we derive some results on birth and death numbers of the population processes $\tilde{N}$ and $\tilde{\tilde{N}}$, needed in Sections 6 and 7 to prove Propositions 1, 2 and 3.

5.1. **Coupling with supercritical birth and death processes during the first phase.** We are interested in the dynamics of the process $\tilde{N}_a$ during the first phase, that is, before the time $\tilde{T}_\varepsilon^K$. The idea is to couple this process with two supercritical birth and death processes, and deduce its dynamics from well known results on birth and death processes. Recall the definition of the rescaled invasion fitness s in (1.10), and for $\varepsilon<S_{aA}/(2C_{a,A}C_{A,a}/C_{A,A}+C_{a,a})$ define the two approximations,

$$s-\frac{2C_{a,A}C_{A,a}+C_{a,a}C_{A,A}}{f_aC_{A,A}}\varepsilon=:s_-(\varepsilon)\leq s\leq s_+(\varepsilon):=s+2\frac{C_{a,A}C_{A,a}}{f_aC_{A,A}}\varepsilon. \tag{5.1}$$

Then for $t<\tilde{T}_\varepsilon^K\wedge S_\varepsilon^K$ the death rate of a-individuals in the process $\tilde{N}$ equals that of the process N, defined in (1.4) and satisfies

$$1-s_+(\varepsilon)\leq\frac{d_a(\tilde{N}(t))}{f_a\tilde{N}_a(t)}=1-s+\frac{C_{a,A}}{f_aK}(\tilde{N}_A(t)-\bar{n}_AK)+\frac{C_{a,a}}{f_aK}\tilde{N}_a(t)\leq 1-s_-(\varepsilon). \tag{5.2}$$

For $S_\varepsilon^K\leq t<\tilde{T}_\varepsilon^K$, according to the definition of $\tilde{N}$ in (3.17), the death rate of a-individuals also satisfies

$$1-s_+(\varepsilon)\leq\frac{d_a^K(\tilde{N}_A(S_\varepsilon^{K-})e_{A\underline{\mathfrak{bc}}},\tilde{N}^{(a)}(t))}{f_a\tilde{N}_a(t)}\leq 1-s_-(\varepsilon). \tag{5.3}$$

Hence, following Theorem 2 in [3] we can construct the processes Z_ε^-, $(\tilde{N}_A, \tilde{N}_a)$ and Z_ε^+ on the same probability space such that almost surely:

$$Z_\varepsilon^-(t) \leq \tilde{N}_a(t) \leq Z_\varepsilon^+(t), \quad \text{for all } t < \tilde{T}_\varepsilon^K, \tag{5.4}$$

where for $* \in \{-,+\}$, Z_ε^* is a birth and death process with initial state 1, and individual birth and death rates f_a and $f_a(1-s_*(\varepsilon))$.

Let σ_u^K denote the time of the first hitting of $\lfloor u \rfloor$ by the process $\tilde{N}_a$:

$$\sigma_u^K := \inf\{t \geq 0, \tilde{N}_a(t) = \lfloor u \rfloor\}, \quad u \in \mathbb{R}_+. \tag{5.5}$$

If for $-1 < s < 1$, $\tilde{Z}^{(s)}$ is a random walk with jumps ± 1 where up-jumps occur with probability $1/(2-s)$ and down-jumps with probability $(1-s)/(2-s)$, we introduce

$$\mathcal{P}_i^{(s)} := \mathcal{L}\left(\tilde{Z}^{(s)} | \tilde{Z}^{(s)}(0) = i\right), \quad i \in \mathbb{N}. \tag{5.6}$$

the law of $\tilde{Z}^{(s)}$ when the initial state is $i \in \mathbb{N}$ and for every $\rho \in \mathbb{R}_+$ the stopping time

$$\tau_\rho := \inf\{n \in \mathbb{Z}_+, \tilde{Z}_n^{(s)} = \lfloor \rho \rfloor\}. \tag{5.7}$$

5.2. Number of jumps of $\tilde{N}_a$ during the first phase.

5.2.1. *Expectation of the number of upcrossings.* Let us recall Equation (4.1) and consider $k < \lfloor \varepsilon K \rfloor$. Then the number of upcrossings from k to $k+1$ during the first phase is:

$$U_k^K(1) := \#\{m, \tau_m^K < \tilde{T}_\varepsilon^K, (\tilde{N}_a(\tau_m^K), \tilde{N}_a(\tau_{m+1}^K)) = (k, k+1)\}, \tag{5.8}$$

where (1) stands for the first phase. Recall (3.1) and (5.1), and introduce a real number λ_ε

$$\lambda_\varepsilon := (1-s_-(\varepsilon))^3(1-s_+(\varepsilon))^{-2}, \tag{5.9}$$

which belongs to $(0,1)$ for ε small enough. We have the following result:

Lemma 5.1. *There exist three positive finite constants c, K_0 and ε_0 such that for $K \geq K_0$ and $\varepsilon \leq \varepsilon_0$:*

If $j \leq k < \lfloor \varepsilon K \rfloor$ and $n_A \in I_\varepsilon^K \pm 1$,

$$\left| \mathbb{E}_{(n_A,j)}^{(1)}[U_k^K(1)] - \frac{1-(1-s)^{\lfloor \varepsilon K \rfloor - k} - (1-s)^{k+1}}{s} \right| \leq c\varepsilon. \tag{5.10}$$

If $k < j < \lfloor \varepsilon K \rfloor$ and $n_A \in I_\varepsilon^K \pm 1$,

$$\mathbb{E}_{(n_A,j)}^{(1)}[U_k^K(1)] \leq \frac{(1-s_-(\varepsilon))^{j-k}}{s_+(\varepsilon)s_-^2(\varepsilon)}. \tag{5.11}$$

If $k' \leq k < \lfloor \varepsilon K \rfloor$ and $n_A \in I_\varepsilon^K \pm 1$,

$$\left| \mathrm{Cov}_{(n_A,j)}^{(1)}(U_k^K(1), U_{k'}^K(1)) \right| \leq c\Big(\lambda_\varepsilon^{(k-k')/2} + \varepsilon\Big). \tag{5.12}$$

Proof. The idea, which comes from [18] and will be used several times throughout Section 5, is to compare the number of upcrossings with geometric random variables. Suppose first that $j \leq k$. Then on the event $\{\tilde{T}_\varepsilon^K < \infty\}$ the process $\tilde{N}_a$ necessarily jumps from k to $k+1$. Being in $k+1$, it either reaches $\lfloor \varepsilon K \rfloor$ before k, or it goes back and then again from k to $k+1$ and so on. We first approximate the probability that there is only one jump from k to $k+1$. As we do not know the value of $\tilde{N}_A$ when $\tilde{N}_a$ hits k for the first time, we bound the probability using the extreme values it can take. Recall Definitions (5.5) and (5.7). The upper bound is derived as follows:

$$\begin{aligned} \mathbb{P}_{(n_A,j)}^{(1)}(U_k^K(1) = 1) \quad &\leq \quad \sup_{n_A \in I_\varepsilon^K \pm 1} \mathbb{P}_{(n_A,k+1)}^{(1)}(\tilde{T}_\varepsilon^K < \sigma_k^K) \\ &= \quad \sup_{n_A \in I_\varepsilon^K \pm 1} \frac{\mathbb{P}_{(n_A,k+1)}(\tilde{T}_\varepsilon^K < \sigma_k^K)}{\mathbb{P}_{(n_A,k+1)}(\tilde{T}_\varepsilon^K < \infty)} \leq q_k^{(s_+(\varepsilon),s_-(\varepsilon))}, \end{aligned} \tag{5.13}$$

where we use (4.3) and for $(s_1, s_2) \in (0,1)^2$

$$q_k^{(s_1,s_2)} := \frac{\mathcal{P}_{k+1}^{(s_1)}(\tau_{\varepsilon K} < \tau_k)}{\mathcal{P}_{k+1}^{(s_2)}(\tau_{\varepsilon K} < \tau_0)}. \tag{5.14}$$

Similarly, we show that $\mathbb{P}^{(1)}_{(n_A,j)}(U_k^K(1)=1)\geq q_k^{(s_-(\varepsilon),s_+(\varepsilon))}$. In the same way, we can approximate the probability that there are least three jumps from k to $k+1$ knowing that there are at least two jumps, and so on. We deduce that we can construct two geometric random variables G_1 and G_2, possibly on an enlarged space, with respective parameters $q_k^{(s_+(\varepsilon),s_-(\varepsilon))}\wedge 1$ and $q_k^{(s_-(\varepsilon),s_+(\varepsilon))}$ such that

$$G_1\leq U_k^K(1)\leq G_2,\quad \text{a.s.} \tag{5.15}$$

In particular, taking the expectation we get from (B.1)

$$\begin{aligned}\frac{(1-(1-s_+(\varepsilon))^{\lfloor\varepsilon K\rfloor-k})(1-(1-s_-(\varepsilon))^{k+1})}{s_+(\varepsilon)(1-(1-s_-(\varepsilon))^{\lfloor\varepsilon K\rfloor})}&\leq \mathbb{E}^{(1)}_{(n_A,j)}[U_k^K(1)]\\ &\leq \frac{(1-(1-s_-(\varepsilon))^{\lfloor\varepsilon K\rfloor-k})(1-(1-s_+(\varepsilon))^{k+1})}{s_-(\varepsilon)(1-(1-s_+(\varepsilon))^{\lfloor\varepsilon K\rfloor})}.\end{aligned} \tag{5.16}$$

According to (1.10) and (5.1), $0<s<1$ and $|s_+(\varepsilon)-s_-(\varepsilon)|\leq(4C_{a,A}C_{A,a}+C_{a,a}C_{A,A})\varepsilon/(f_aC_{A,A})$. Hence the last inequality and straightforward calculations lead to (5.10).

Let us now assume that $k<j$. Then we have

$$\begin{aligned}\mathbb{P}^{(1)}_{(n_A,j)}(U_k^K(1)\geq 1)\leq \sup_{n_A\in I_\varepsilon^K\pm 1}\mathbb{P}^{(1)}_{(n_A,j)}(\sigma_k^K<\tilde{T}_\varepsilon^K)&=\sup_{n_A\in I_\varepsilon^K\pm 1}\frac{\mathbb{P}_{(n_A,j)}(\tilde{T}_\varepsilon^K<\infty|\sigma_k^K<\tilde{T}_\varepsilon^K)\mathbb{P}_{(n_A,j)}(\sigma_k^K<\tilde{T}_\varepsilon^K)}{\mathbb{P}_{(n_A,j)}(\tilde{T}_\varepsilon^K<\infty)}\\ &\leq\frac{\mathcal{P}_k^{(s_+(\varepsilon))}(\tau_{\varepsilon K}<\tau_0)\mathcal{P}_j^{(s_-(\varepsilon))}(\tau_k<\tau_{\varepsilon K})}{\mathcal{P}_j^{(s_-(\varepsilon))}(\tau_{\varepsilon K}<\tau_0)}\leq\frac{(1-s_-(\varepsilon))^{j-k}}{s_+(\varepsilon)s_-(\varepsilon)},\end{aligned}$$

where we again used (4.3) and (B.1). Moreover, the same proof as for (5.15) leads to:

$$\mathbb{E}^{(1)}_{(n_A,j)}[U_k^K(1)|U_k^K(1)\geq 1]\leq\left(q_k^{(s_-(\varepsilon),s_+(\varepsilon))}\right)^{-1}\leq s_-^{-1}(\varepsilon),$$

where we used Equation (B.3). This ends the proof of (5.11). The last inequality, (5.12), has been stated in [19] (Equation (7.26)). □

5.2.2. *Expectation of hitting numbers.* Let us recall (5.8) and introduce for $0<j\leq k<\lfloor\varepsilon K\rfloor$ the total number of downcrossings from k to $k-1$,

$$D_k^K(1):=\#\{m,\tau_m^K\leq\tilde{T}_\varepsilon^K,(\tilde{N}_a(\tau_m^K),\tilde{N}_a(\tau_{m+1}^K))=(k,k-1)\}, \tag{5.17}$$

and the number of hittings of the state k by the process $\tilde{N}_a$ before the time $\tilde{T}_\varepsilon^K$:

$$V_k^K(1):=U_{k-1}^K(1)+D_{k+1}^K(1)=\#\{m,\tau_m^K\leq\tilde{T}_\varepsilon^K,\tilde{N}_a(\tau_{m-1}^K)\neq k,\tilde{N}_a(\tau_m^K)=k\}. \tag{5.18}$$

Recall the definition of $\lambda_\varepsilon\in(0,1)$ in (5.9). We can state the following Lemma, which will be useful to get bounds on the number of upcrossings of the A-population during the first phase (see Lemma 5.4):

Lemma 5.2. *There exist three finite constants c, K_0 and ε_0 such that for $K\geq K_0$, $\varepsilon\leq\varepsilon_0$ and $k'<k<\lfloor\varepsilon K\rfloor$:*

$$\left|\mathbb{E}^{(1)}[V_k^K(1)]-\frac{(2-s)(1-(1-s)^{\lfloor\varepsilon K\rfloor-k}-(1-s)^k)}{s}\right|\leq c\varepsilon,\quad \textit{and}\quad |\operatorname{Cov}^{(1)}(V_{k'}^K(1),V_k^K(1))|\leq c(\varepsilon+\lambda_\varepsilon^{(k-k')/2}).$$

Proof. Under $\mathbb{P}^{(1)}$ the a-population size goes from 1 to $\lfloor\varepsilon K\rfloor$, thus the number of downcrossings from $k+1$ to k is equal to the number of upcrossings from k to $k+1$ minus 1. Adding (5.18) yields

$$V_k^K(1)=U_{k-1}^K(1)+U_k^K(1)-1,\quad \mathbb{P}^{(1)}-a.s.$$

We get the first part of the Lemma by taking the expectation and applying (5.10). The proof of the second part follows that of (5.12), and once again we can find the details in the proof of Equation (7.26) in [19]. □

5.2.3. *Number of upcrossings during an excursion above or below a given level.* We now focus on the number of upcrossings from k to $k+1$ during an excursion above or below l. Let us denote by $\sigma_l^K(1)$ the jump number of the first hitting of l before the end of the first phase: for $l < \lfloor \varepsilon K \rfloor$,

$$\sigma_l^K(1) := \inf\{m, \tau_m^K \leq \tilde{T}_\varepsilon^K, \tilde{N}_a(\tau_m^K) = l\}, \tag{5.19}$$

and for $1 \leq k, l < \lfloor \varepsilon K \rfloor$ and $n_A \in I_\varepsilon^K \pm 1$,

$$U_{n_A,l,k}^K(1) := \#\Big\{m < \sigma_l^K(1), (\tilde{N}_a(\tau_m^K), \tilde{N}_a(\tau_{m+1}^K)) = (k, k+1)\Big\}. \tag{5.20}$$

Then, if we denote by μ_ε the real number

$$\mu_\varepsilon := (1 - s_-(\varepsilon))^2 (1 - s_+(\varepsilon))^{-1}, \tag{5.21}$$

which belongs to $(0,1)$ for ε small enough, we can derive the following bounds:

Lemma 5.3. *There exist three positive, finite constants c, K_0 and ε_0 such that for $K \geq K_0$, $\varepsilon \leq \varepsilon_0$, $1 \leq k < l < \lfloor \varepsilon K \rfloor$ and $n_A \in I_\varepsilon^K \pm 1$,*

$$\mathbb{E}^{(1)}_{(n_A,k+1)}[U_{n_A,k,l}^K(1) | \sigma_k^K(1) < \infty] \vee \mathbb{E}^{(1)}_{(n_A,l-1)}[U_{n_A,l,k}^K(1)] \leq c\mu_\varepsilon^{l-k}.$$

Proof. Equations (B.5) and (B.6) in [19] state that for $k < l < \lfloor \varepsilon K \rfloor$ and $n_A \in I_\varepsilon^K \pm 1$,

$$\mathbb{P}^{(1)}_{(n_A,k+1)}(U_{n_A,k,l}^K(1) \geq 1 | \sigma_k^K(1) < \infty) \leq c(1 - s_-(\varepsilon))^{l-k},$$

and

$$\mathbb{P}^{(1)}_{(n_A,k+1)}(U_{n_A,k,l}^K(1) = 1 | U_{n_A,k,l}^K(1) \geq 1, \sigma_k^K(1) < \infty) \geq c\Big(\frac{1 - s_+(\varepsilon)}{1 - s_-(\varepsilon)}\Big)^{l-k}$$

for a finite c. By comparing $U_{n_A,k,l}^K(1)$ with a geometric random variable we get the first inequality. To bound the expectation of upcrossings from k to $k+1$ during an excursion below l we first bound the probability to have at least one jump from k to $k+1$ during such an excursion. By definition, $\tilde{N}_a$ necessarily hits $l-1$ during the excursion below l. Recall Definitions (4.3), (5.5) and (5.7). Then for every n_A in $I_\varepsilon^K \pm 1$,

$$\begin{aligned}
\mathbb{P}^{(1)}_{(n_A,l-1)}(\sigma_k^K < \sigma_l^K | \sigma_l^K < \infty) &= \frac{\mathbb{P}_{(n_A,l-1)}(\tilde{T}_\varepsilon^K < \infty | \sigma_k^K < \sigma_l^K)\mathbb{P}_{(n_A,l-1)}(\sigma_k^K < \sigma_l^K)}{\mathbb{P}_{(n_A,l-1)}(\tilde{T}_\varepsilon^K < \infty)} \\
&\leq \frac{\mathcal{P}_k^{(s_+(\varepsilon))}(\tau_{\varepsilon K} < \tau_0)\mathcal{P}_{l-1}^{(s_-(\varepsilon))}(\tau_k < \tau_l)}{\mathcal{P}_{l-1}^{(s_-(\varepsilon))}(\tau_{\varepsilon K} < \tau_0)} \leq \frac{(1 - s_-(\varepsilon))^{l-k-1}}{s_-(\varepsilon)},
\end{aligned}$$

where we used (B.1). The next step consists in bounding the number of upcrossings from k to $k+1$ during the excursion knowing that this number is larger than one: for $n_A \in I_\varepsilon^K \pm 1$,

$$\begin{aligned}
\mathbb{P}^{(1)}_{(n_A,k+1)}(\sigma_l^K < \sigma_k^K) &= \frac{\mathbb{P}_{(n_A,k+1)}(\tilde{T}_\varepsilon^K < \infty | \sigma_l^K < \sigma_k^K)\mathbb{P}_{(n_A,k+1)}(\sigma_l^K < \sigma_k^K)}{\mathbb{P}_{(n_A,k+1)}(\tilde{T}_\varepsilon^K < \infty)} \\
&\geq \frac{\mathcal{P}_l^{(s_-(\varepsilon))}(\tau_{\varepsilon K} < \tau_0)\mathcal{P}_{k+1}^{(s_-(\varepsilon))}(\tau_l < \tau_k)}{\mathcal{P}_{k+1}^{(s_+(\varepsilon))}(\tau_{\varepsilon K} < \tau_0)} \geq s_-^2(\varepsilon),
\end{aligned}$$

where we again used (B.1). Hence on the event $\{U_{n_A,l,k}^K(1) \geq 1\}$, $U_{n_A,l,k}^K(1)$ is smaller than a geometric random variable with parameter $s_-^2(\varepsilon)$ and we get:

$$\mathbb{E}^{(1)}_{(n_A,l-1)}[U_{n_A,l,k}^K(1)] \leq s_-^{-2}(\varepsilon)\mathbb{P}^{(1)}_{(n_A,l-1)}(U_{n_A,l,k}^K(1) \geq 1) \leq \frac{(1 - s_-(\varepsilon))^{l-k-1}}{s_-^3(\varepsilon)},$$

which ends the proof of Lemma 5.3. $\square$

5.3. **Number of jumps $\tilde{N}_A$ during the first phase.** We introduce for $k < \lfloor \varepsilon K \rfloor$ the number of upcrossings of the A-population when the a-population is of size k:

$$\mathcal{U}_k^K(1) := \#\{m, \tau_m^K \leq \tilde{T}_\varepsilon^K, \tilde{N}_A(\tau_{m+1}^K) - \tilde{N}_A(\tau_m^K) = 1, \tilde{N}_a(\tau_m^K) = k\}. \tag{5.22}$$

We are now able to get bounds for the expectations and covariances of these quantities:

Lemma 5.4. *There exist three finite constants c, K_0 and ε_0 such that for $K \geq K_0$, $\varepsilon \leq \varepsilon_0$ and $k < \lfloor \varepsilon K \rfloor$,*

$$\Big| \mathbb{E}^{(1)} \Big[\sum_{i=1}^k \mathcal{U}_i^K(1) \Big] - \frac{f_A \bar{n}_A K \log k}{s f_a} \Big| \leq cK(1 + \varepsilon \log k) \quad \textit{and} \quad \mathrm{Var}^{(1)} \Big(\sum_{i=1}^k \mathcal{U}_i^K(1) \Big) \leq cK^2(1 + \varepsilon \log^2 k).$$

Proof. The proof is based on the comparison of the A- and a-population jump rates. Let us first focus on the a-population. For $k \leq \lfloor \varepsilon K \rfloor$ and $n_A \in I_\varepsilon^K \pm 1$,

$$\begin{aligned}
\mathbb{P}^{(1)}_{(n_A,k)}(\tilde{N}_a(\delta t) \neq k) &= \mathbb{P}_{(n_A,k)}(\tilde{T}_\varepsilon^K < \infty | \tilde{N}_a(\delta t) = k \pm 1) \mathbb{P}_{(n_A,k)}(\tilde{N}_a(\delta t) = k \pm 1) / \mathbb{P}_{(n_A,k)}(\tilde{T}_\varepsilon^K < \infty) \\
&\leq \mathcal{P}_{k \pm 1}^{(s_+(\varepsilon))}(\tilde{T}_\varepsilon^K < \infty) \mathbb{P}_{(n_A,k)}(\tilde{N}_a(\delta t) = k \pm 1) / \mathcal{P}_k^{(s_-(\varepsilon))}(\tilde{T}_\varepsilon^K < \infty) \\
&\leq \frac{(1 + c\varepsilon)}{1 - (1-s)^k} \Big((1 - (1-s)^{k+1}) f_a k + (1-s)(1 - (1-s)^{k-1})(D_a + C_{aA} \bar{n}_A) k \Big) \delta t \\
&= (1 + c\varepsilon) f_a (2 - s) k \delta t,
\end{aligned} \tag{5.23}$$

for a finite constant c and ε small enough, where δt is a small time step and by abuse of notation we did not indicate the $o(\delta t)$'s. We used the definition of $\mathbb{P}^{(1)}$ in (4.3) for the equality, Coupling (5.4) for the first inequality, (B.1) for the second one, and the equality $S_{aA} = f_a - D_a - C_{a,A} \bar{n}_A$ for the last one. Reasoning similarly we get:

$$(1 - c\varepsilon) f_a (2 - s) k \delta t \leq \mathbb{P}^{(1)}_{(n_A,k)}(\tilde{N}_a(\delta t) \neq k). \tag{5.24}$$

Let us now focus on the number of upcrossings of the A-population. The definition of $\tilde{N}$ in (3.17) and Bayes' Theorem yield

$$(1 - c\varepsilon) f_A \bar{n}_A K \delta t \leq \mathbb{P}^{(1)}_{(n_A,k)}(\tilde{N}_A(\delta t) = n_A + 1) \leq (1 + c\varepsilon) f_A \bar{n}_A K \delta t, \tag{5.25}$$

for a finite c and ε small enough. Indeed, from Coupling (5.4) and Equation (B.1) we get the following bound, independent of n_A in $I_\varepsilon^K \pm 1$:

$$\frac{1 - (1 - s_-(\varepsilon))^k}{1 - (1 - s_-(\varepsilon))^{\lfloor \varepsilon K \rfloor}} \leq \mathbb{P}_{(n_A,k)}(\tilde{T}_\varepsilon^K < \infty) \leq \frac{1 - (1 - s_+(\varepsilon))^k}{1 - (1 - s_+(\varepsilon))^{\lfloor \varepsilon K \rfloor}}.$$

Hence there exist two finite constants c and ε_0 such that for every $\varepsilon \leq \varepsilon_0$, if we introduce the parameters

$$\frac{1}{q_k^{(1)}(\varepsilon)} := 1 + (1 - c\varepsilon) \frac{f_A \bar{n}_A K}{(2 - s) f_a k} \leq 1 + (1 + c\varepsilon) \frac{f_A \bar{n}_A K}{(2 - s) f_a k} =: \frac{1}{q_k^{(2)}(\varepsilon)}, \tag{5.26}$$

we can deduce from (5.23) to (5.25) that for $k < \lfloor \varepsilon K \rfloor$

$$\sum_{V_k^K(1)} \Big(G^i_{q_k^{(1)}(\varepsilon)} - 1 \Big) \leq \mathcal{U}_k^K(1) \leq \sum_{V_k^K(1)} \Big(G^i_{q_k^{(2)}(\varepsilon)} - 1 \Big), \tag{5.27}$$

where for $j \in \{1, 2\}$, $(G^i_{q_k^{(j)}(\varepsilon)}, i \in \mathbb{N})$ is a sequence of geometric random variables with parameter $q_k^{(j)}(\varepsilon)$ independent of $V_l^K(1)$ (defined in (5.18)) for all $l < \lfloor \varepsilon K \rfloor$. Hence a direct application of Lemmas 5.2 and B.2 leads to

$$\Big| \mathbb{E}^{(1)} \Big[\mathcal{U}_k^K(1) \Big] - \frac{f_A \bar{n}_A K}{s f_a k} (1 - (1-s)^k - (1-s)^{\lfloor \varepsilon K \rfloor - k}) \Big| \leq c\varepsilon \frac{K}{k}, \tag{5.28}$$

for a finite c and ε small enough. This implies the first inequality of Lemma 5.4.

Let us now bound the second moment of $\mathcal{U}_k^K(1)$ and the expectation of $\mathcal{U}_k^K(1) \mathcal{U}_l^K(1)$ for $k \neq l$. The first upper bound follows again from a direct application of Lemmas 5.2 and B.2. We get

$$\mathbb{E}^{(1)} \Big[(\mathcal{U}_k^K(1))^2 \Big] \leq \mathbb{E}^{(1)} \Big[\Big(\sum_{V_k^K(1)} G^i_{q_k^{(2)}(\varepsilon)} \Big)^2 \Big] \leq \frac{2(\mathbb{E}^{(1)}[V_k^K(1)])^2}{(q_k^{(2)}(\varepsilon))^2} \leq 2(1 + c\varepsilon) \Big(\frac{f_A \bar{n}_A K}{s f_a k} \Big)^2, \tag{5.29}$$

for a finite c and ε small enough. A new application of the same Lemmas yields, for $k < l < \lfloor \varepsilon K \rfloor$

$$\mathbb{E}^{(1)}\Big[\mathcal{U}_k^K(1)\mathcal{U}_l^K(1)\Big] \leq \frac{\mathbb{E}^{(1)}[V_k^K(1)V_l^K(1)]}{q_k^{(2)}(\varepsilon)q_l^{(2)}(\varepsilon)} \leq c(1+\varepsilon+\lambda_\varepsilon^{(l-k)/2})\frac{(f_A\bar{n}_AK)^2}{(f_as)^2kl}, \tag{5.30}$$

where we used that $\mathbb{E}^{(1)}[XY] = \mathbb{E}^{(1)}[X]\mathbb{E}^{(1)}[Y] + \mathrm{Cov}^{(1)}(X,Y)$ for any real random variables (X,Y). From (5.27) to (5.30) and (B.2) we deduce that there exists a finite c such that for ε small enough and $k < \lfloor \varepsilon K \rfloor$,

$$\mathbb{E}^{(1)}\Big[\Big(\sum_{i=1}^k \mathcal{U}_i^K(1)\Big)^2\Big] \leq (1+c\varepsilon)\Big(\frac{f_A\bar{n}_AK\log k}{f_as}\Big)^2 + cK^2. \tag{5.31}$$

Reasoning similarly to get the lower bound, we obtain

$$\Big|\mathbb{E}^{(1)}\Big[\Big(\sum_{i=1}^k \mathcal{U}_i^K(1)\Big)^2\Big] - \Big(\frac{f_A\bar{n}_AK\log k}{f_as}\Big)^2\Big| \leq cK^2(1+\varepsilon\log^2 k). \tag{5.32}$$

Adding the first inequality of Lemma 5.4 we conclude the proof. □

5.4. **Coupling with subcritical birth and death processes during the third phase.** We couple the process $\tilde{\tilde{N}}_a$ with two subcritical birth and death processes to control its dynamics. We recall the definition of $\mathcal{N}_\varepsilon^K$ in (3.12) and introduce

$$\bar{s} := |S_{Aa}|/f_A. \tag{5.33}$$

Let us define for ε small enough,

$$\bar{s} - \frac{M''C_{A,a}}{f_A}\varepsilon =: \bar{s}_-(\varepsilon) < \bar{s} < \bar{s}_+(\varepsilon) := \bar{s} + \frac{C_{A,A}+M''C_{A,a}}{f_A}\varepsilon, \tag{5.34}$$

where M'' has been defined just before Definition (3.10). Then, according to the definition of $\tilde{\tilde{N}}$ in (3.18), we can follow Theorem 2 in [3] and construct the processes Y_ε^+, $\tilde{\tilde{N}}$ and Y_ε^- on the same probability space such that on the event $\mathcal{N}_\varepsilon^K$

$$Y_\varepsilon^+(t) \leq \tilde{\tilde{N}}_A(t) \leq Y_\varepsilon^-(t), \quad \text{for all } T_\varepsilon^K + t_\varepsilon \leq t < T_\varepsilon^K + t_\varepsilon + \tilde{\tilde{T}}_0^{(K,A)}, \quad \text{a.s.}, \tag{5.35}$$

where for $* \in \{-,+\}$, Y_ε^* is a birth and death process with initial state $N_A(T_\varepsilon^K + t_\varepsilon)$ and individual birth and death rates f_A and $f_A(1+\bar{s}_*(\varepsilon))$, and we recall that $\tilde{\tilde{T}}_0^{(K,A)}$ is the analog of $T_0^{(K,A)}$ (defined in (3.11)) for the process $\tilde{\tilde{N}}$.

Recall Definition (5.6), and let us introduce for $i \in \mathbb{N}$ $\mathcal{Q}_i^{(s)} = \mathcal{P}_i^{(-s)}$ and for $\rho \in \mathbb{R}_+$ the stopping time

$$\nu_\rho := \inf\{n \in \mathbb{Z}_+, \tilde{Z}_n^{(-s)} = \lfloor \rho \rfloor\}. \tag{5.36}$$

5.5. **Number of jumps of $\tilde{\tilde{N}}_A$ during the third phase.** Similarly as in (5.18) we introduce for $1 \leq k < \lfloor \varepsilon K \rfloor$ the random variable $\mathcal{V}_k^K(3)$ which corresponds to the number of hittings of state k by the process $\tilde{\tilde{N}}_A$ during the third phase. Recall Definitions (3.9), (3.10) and (5.34). We have the following approximations:

Lemma 5.5. *Let u be in $[\omega_1,\omega_2]$. There exist three finite constants c, K_0 and ε_0 such that for $K \geq K_0$, $\varepsilon \leq \varepsilon_0$ and n_a in $J_\varepsilon^K \pm 1$, if $\lfloor uK \rfloor < k < \lfloor \varepsilon K \rfloor$,*

$$\mathbb{E}^{(3)}_{(\lfloor uK \rfloor, n_a)}[\mathcal{V}_k^K(3)] \leq (1+c\varepsilon)\frac{2+\bar{s}}{\bar{s}}(1+\bar{s}_-(\varepsilon))^{\lfloor uK \rfloor - k},$$

and if $k \leq \lfloor uK \rfloor$,

$$\Big|\mathbb{E}^{(3)}_{(\lfloor uK \rfloor, n_a)}[\mathcal{V}_k^K(3)] - \frac{2+\bar{s}}{\bar{s}}(1-(1+\bar{s})^{-k} - (1+\bar{s})^{k-\lfloor \varepsilon K \rfloor})\Big| \leq c\varepsilon.$$

Proof. The proof is very similar to that of (5.10), hence we do not detail all the calculations and refer to the proof of Lemma 5.1. First we consider $\lfloor uK \rfloor < k < \lfloor \varepsilon K \rfloor$ and approximate under $\mathbb{P}^{(3)}$ the probability for $\tilde{\tilde{N}}_A$ to hit k before the extinction of the A-population. Indeed, if $k \leq \lfloor uK \rfloor$, we know that $\tilde{\tilde{N}}_A$ hits k $\mathbb{P}^{(3)}$-a.s. Let $\lfloor uK \rfloor < k < \lfloor \varepsilon K \rfloor$. Then for every $n_a \in J_\varepsilon^K \pm 1$, Equation (B.1) implies

$$\mathbb{P}^{(3)}_{(\lfloor uK \rfloor, n_a)}(\tilde{\tilde{N}}_A \text{ hits } k) \leq \frac{\mathcal{Q}_k^{(\bar{s}_+(\varepsilon))}(\nu_0 < \nu_{\varepsilon K})\mathcal{Q}_{\lfloor uK \rfloor}^{(\bar{s}_-(\varepsilon))}(\nu_k < \nu_0)}{\mathcal{Q}_{\lfloor uK \rfloor}^{(\bar{s}_-(\varepsilon))}(\nu_0 < \nu_{\varepsilon K})} \leq \frac{1+c\varepsilon}{(1+\bar{s}_-(\varepsilon))^{k-\lfloor uK \rfloor}}, \tag{5.37}$$

for a finite c, ε small enough and K large enough. The second step consists in counting how many times the process $\tilde{\tilde{N}}_A$ hits k during the third phase knowing that it happens at least once. Once again we will compare this number with geometric random variables, by approximating the probability to have only one jump. The following inequality follows the spirit of (5.13). The only difference is that in the third phase $\tilde{\tilde{N}}_A$ is coupled with subcritical birth and death processes, whereas in the first phase $\tilde{N}_a$ was coupled with supercritical birth and death processes. For every $n_a \in J_\varepsilon^K \pm 1$ and $k < \lfloor \varepsilon K \rfloor$,

$$\mathbb{P}^{(3)}_{(k,n_a)}(\tilde{\tilde{N}}_A(t) \leq k, \forall t \geq 0) \geq \frac{\mathcal{Q}_{k-1}^{(\bar{s}_-(\varepsilon))}(\nu_0 < \nu_k)\mathcal{Q}_k^{(\bar{s}_-(\varepsilon))}(\nu_{k-1} < \nu_{k+1})}{\mathcal{Q}_k^{(\bar{s}_+(\varepsilon))}(\nu_0 < \nu_{\varepsilon K})} \geq \frac{(1-c\varepsilon)\bar{s}}{(2+\bar{s})(1-(1+\bar{s})^{-k}-(1+\bar{s})^{k-\lfloor \varepsilon K \rfloor})}.$$

We derive the upper bound similarly and end the proof by comparing the hitting numbers with geometric random variables. For $\lfloor uK \rfloor < k < \lfloor \varepsilon K \rfloor$ we have to multiply the expectation of the geometric random variables by the probability to hit k at least once, approximated in (5.37). □

5.6. **Number of births of a-individuals during the third phase.** Recall (3.16) and let $U_k^K(3)$ be the number of births in the a-population during the third phase when $\tilde{\tilde{N}}_A$ equals $k \leq \lfloor \varepsilon K \rfloor$

$$U_k^K(3) := \#\{m, T_\varepsilon^K + t_\varepsilon < \tau_m^K \leq T_{\text{ext}}^K, \tilde{\tilde{N}}_A(\tau_m^K) = k, \text{ and } \{\{\tilde{\tilde{N}}_a(\tau_{m+1}^K) - \tilde{\tilde{N}}_a(\tau_m^K) = 1\} \\ \text{or } \{\tilde{\tilde{N}}_a(\tau_{m+1}^K) = \tilde{\tilde{N}}_a(\tau_m^K), \tilde{\tilde{N}}^{(a)}(\tau_{m+1}^K) \neq \tilde{\tilde{N}}^{(a)}(\tau_m^K)\}\}. \tag{5.38}$$

We now state an approximation for the expectation of $U_k^K(3)$. We do not prove this result as it is obtained in the same way as Lemma 5.4: the birth rate of the a-population is close to $f_a\bar{n}_aK$, the jump rate of the A-population is of order $(2+\bar{s})f_Ak$ when $\tilde{\tilde{N}}_A = k$ and the expectations of the hitting numbers for the A-population are given in Lemma 5.5. The only difference is that the A-population size can hit values bigger than the initial value of the third phase, $\tilde{\tilde{N}}_A(T_\varepsilon^K + t_\varepsilon)$. However the probabilities to hit such values decrease geometrically (see Lemma 5.5) and they have a negligible influence on the final result. Thus we get

Lemma 5.6. *There exist three finite constants c, ε_0 and K_0 such that for $\varepsilon \leq \varepsilon_0$, $K \geq K_0$ and $k \leq \lfloor \varepsilon K \rfloor$*

$$\Big|\mathbb{E}^{(3)}\Big[\sum_{i=1}^k U_i^K(3)\Big] - \frac{f_a\bar{n}_aK\log k}{\bar{s}f_A}\Big| \leq cK(1+\varepsilon\log k) \quad \textit{and} \quad \text{Var}^{(3)}\Big(\sum_{i=1}^k U_i^K(3)\Big) \leq cK^2(1+\varepsilon\log^2 K).$$

6. First phase

This section is dedicated to the proof of Proposition 1. We prove that there are only four different possible ancestral relationships of the two neutral loci and calculate the probabilities for the non-negligible possibilities.

6.1. **Coalescence and recombination probabilities, negligible events.** Recall Definition 4.1 and define, for $j \in \{1,2\}$

$$r_j^* := r_1 + \mathbf{1}_{\{j=2\}}(r_2 - 2r_1r_2), \quad \text{and} \quad r_{(1,2)}^* := r_1r_2,$$

which denote the probability to have one (resp. two) recombination(s) somewhere before the locus Nj (resp. before the locus $N2$) at a birth event.

Definition 6.1. *For $(\alpha,\alpha') \in \mathcal{A}^2$, $j \in \{1,2\}$ and $n = (n_A, n_a) \in \mathbb{N}^{\mathcal{A}}$ we define:*

$p_{\alpha\alpha'}^{(c,j)}(n)$ *:= probability that two randomly chosen neutral alleles, located at locus Nj and associated respectively with alleles α and α' at time τ_m^K, coalesce at this time conditionally on $(N_A, N_a)(\tau_{m-1}^K) = n$ and on the birth of an individual carrying allele α at time τ_m^K.*

$p^{(j)}_{\alpha\alpha'}(n)$:= *probability to have one (and only one) recombination from the* α- *into the* α'-*population before locus* Nj *conditionally on* $(N_A,N_a)(\tau^K_{m-1})=n$ *and on the birth of an individual carrying allele* α *at time* τ^K_m.

$p^{(1,2)}_{\alpha\alpha'}(n)$:= *probability to have a double recombination under the same conditions*

Then we have the following result:

Lemma 6.1. *Let* $\alpha\in\mathcal{A}$, $n=(n_A,n_a)\in\mathbb{N}^{\mathcal{A}}$ *such that* $n_a\leq\lfloor\varepsilon K\rfloor$, $n_A\in I^K_\varepsilon\pm 1$ *and* $j\in\{1,2\}$. *Then there exists a finite* c *such that,*

$$p^{(c,j)}_{aa}(n)=\frac{2}{n_a(n_a+1)}\Big(1-\frac{r^*_jf_An_A}{f_An_A+f_an_a}\Big),\quad p^{(c,j)}_{aA}(n)=\frac{r^*_jf_A}{(n_a+1)(f_An_A+f_an_a)}\quad \textit{and}\quad p^{(c,j)}_{A\alpha}(n)\leq\frac{c}{K^2}.$$

Proof. The proof of the two equalities can be found in [19] (Lemma 7.1) as the expression is the same for $n_A\in I^K_\varepsilon$ or $dist(n_A,I^K_\varepsilon)=1$ (where $dist$ is the canonical distance on $\mathbb{R}$). The only difference is that we consider two neutral loci and have to exclude the double recombination case. Indeed, if there are simultaneous recombinations the alleles located at SL and N2 in the newborn originate from the same parent. The expressions of $p^{(c,j)}_{A\alpha}(n)$ in the case where $n_A\in I^K_\varepsilon$ are also stated in [19] (Lemma 7.1), and from the definition of $\tilde{N}$ in (3.17) we get that when $dist(n_A,I^K_\varepsilon)=1$, $p^{(c,j)}_{AA}(n)=2/n_A^2$ and $p^{(c,j)}_{Aa}(n)=0$. This ends the proof. □

Next we focus on the recombination probabilities:

Lemma 6.2. *Let* $\alpha\in\mathcal{A}$, $n=(n_A,n_a)\in\mathbb{N}^{\mathcal{A}}$ *such that* $n_a\leq\lfloor\varepsilon K\rfloor$, $n_A\in I^K_\varepsilon\pm 1$ *and* $j\in\{1,2,(1,2)\}$. *Then there exist two finite constants* c *and* ε_0 *such that for every* $\varepsilon\leq\varepsilon_0$,

$$p^{(j)}_{aa}(n)=\frac{r^*_jf_a(n_a-1)}{(n_a+1)(f_An_A+f_an_a)},\quad p^{(j)}_{aA}(n)=\frac{r^*_jf_An_A}{(n_a+1)(f_An_A+f_an_a)},$$

$$p^{(j)}_{Aa}(n)\leq\frac{c\varepsilon}{K\log K}\quad \textit{and}\quad (1-c\varepsilon)\frac{r_2}{n_A}\leq p^{(2)}_{AA}(n_A,k)\leq\frac{r_2}{n_A},\ k\leq\lfloor\varepsilon K\rfloor \tag{6.1}$$

Proof. The second equality is stated in [19] Equation (7.2).

Conditionally on the birth of an a-individual and the state of the process at the $(m-1)$-th jump, the probability of picking the newborn when choosing an individual at random amongst the a-individuals is equal to $1/(n_a+1)$. A recombination before the locus Nj (or before locus $N1$ and locus $N2$ if $j=(1,2)$) happens with probability r^*_j, independent of all other events. Finally, the probability that the second parent is an a-individual but is different from the first parent is equal to $f_a(n_a-1)/(f_An_A+f_an_a)$. This proves the first equality.

When $n_A\in I^K_\varepsilon$ we get similarly that

$$p^{(j)}_{AA}(n)=\frac{r^*_jf_A(n_A-1)}{(n_A+1)(f_An_A+f_an_a)}\quad\text{and}\quad p^{(j)}_{Aa}(n)=\frac{r^*_jf_an_a}{(n_A+1)(f_An_A+f_an_a)},$$

and from the definition of $\tilde{N}$ in (3.17) we obtain that when $dist(n_A,I^K_\varepsilon)=1$, $p^{(2)}_{AA}(n)=r_2(n_A-1)/n_A^2$ and $p^{(j)}_{Aa}(n)=0$. Condition (1.1) completes the proof. □

Remark 1. *Let us recall the definition of* I^K_ε *in* (3.1). *Then there exist three finite constants* c, ε_0 *and* K_0 *such that for* $\varepsilon\leq\varepsilon_0$, $K\geq K_0$, $j\in\{1,2,(1,2)\}$, $n_A\in I^K_\varepsilon\pm 1$ *and* $k<\lfloor\varepsilon K\rfloor$,

$$(1-c\varepsilon)\frac{r^*_j}{k+1}\leq p^{(j)}_{aA}(n_A,k)\leq\frac{r^*_j}{k+1}\quad \textit{and}\quad p^{(2)}_{aa}(n_A,k)\leq\frac{f_a}{f_A}\frac{r_2}{n_A}\leq\frac{c}{K\log K}. \tag{6.2}$$

Recalling the definitions of the mth jump time and the number of jumps in (4.1) and (4.2), we define for $j\in\{1,2,(1,2)\}$, $m\in\mathbb{N}$ and an individual i uniformly picked at the end of the first phase,

$$(\alpha ij)_m:=\{m\leq J^K(1)\text{ and the }j\text{-th locus/loci of the }i\text{-th individual is/are associated to an allele }\alpha\text{ at the }m\text{-th jump time}\}. \tag{6.3}$$

The notation $(\alpha i1)_m,(\alpha' i2)_m$ here implies that the two neutral loci of individual i are associated to two distinct individuals at the mth jump time, for any $\alpha,\alpha'\in\mathcal{A}$.

To approximate the genealogy of the neutral alleles sampled at the end of the first phase we will focus on the recombinations and coalescences which may happen during this time interval. Keep in mind that when looking at coalescing neutral loci, the parent's type may differ from the type of up to one child. We first prove that we can neglect some event combinations. Sample $2d$ distinct individuals uniformly at the end of the first phase (maximal number of ancestors for the $2d$ neutral alleles sampled at the end of the sweep) and define:

- aAa: a neutral allele recombines from the a-population to the A-population, and then (backwards in time) back into the a-population
- CR: two neutral alleles coalesce in the a-population, and then (backwards in time) recombine into the A-population
- CA: two neutral alleles coalesce and at least one of them carries the allele A at the time of coalescence
- $2R$: a neutral allele takes part in a double recombination (i.e. a recombination before $N1$ and a recombination before $N2$ at the same birth event)
- $R2a$: a recombination separates the two neutral loci of an individual within the a-population

We can bound the probability of these events as follows:

Lemma 6.3. *There exist three positive finite constants c, K_0 and ε_0 such that for $\varepsilon \leq \varepsilon_0$ and $K \geq K_0$*

$$\mathbb{P}^{(1)}(aAa) + \mathbb{P}^{(1)}(CR) + \mathbb{P}^{(1)}(2R) + \mathbb{P}^{(1)}(R2a) \leq \frac{c}{\log K}, \quad \textit{and} \quad \mathbb{P}^{(1)}(CA) \leq \frac{c \log K}{K}.$$

Proof. The probabilities of events aAa, CR and CA are bounded in [19] Lemma 7.3 and Equation (7.19) for the process N. But according to Lemmas 6.1 and 6.2 the coalescence and recombination probabilities for the process $\tilde{N}$ are very close or even smaller when $dist(n_A, I_\varepsilon^K) = 1$ than when N and $\tilde{N}$ are equal. Hence we just have to bound the probability of $2R$ and $R2a$. If a neutral allele experiences a double recombination, it happens either when it is associated with an allele a, or with an allele A. From Lemma 6.2 and the fact that r_1 and r_2 are of order $1/\log K$ we get for $k < \lfloor \varepsilon K \rfloor$:

$$\sup_{n_A \in I_\varepsilon^K \pm 1} \Big(p_{aa}^{(1,2)} + p_{aA}^{(1,2)} \Big)(n_A, k) \leq \frac{c}{(k+1)\log^2 K} \quad \text{and} \quad \sup_{n_A \in I_\varepsilon^K \pm 1} \Big(p_{Aa}^{(1,2)} + p_{AA}^{(1,2)} \Big)(n_A, k) \leq \frac{c}{K \log^2 K}.$$

Recall the definitions of $U_k^K(1)$ and $\mathcal{U}_k^K(1)$ in (5.8) and (5.22) respectively. As a birth of an α-individual is needed to have a recombination from the α- to the α'-population, we can bound the probability to have a double recombination by:

$$\mathbb{P}^{(1)}(2R) \leq \frac{c}{\log^2 K} \mathbb{E}^{(1)} \Big[\sum_{k=1}^{\lfloor \varepsilon K \rfloor - 1} \Big(\frac{U_k^K(1)}{k+1} + \frac{\mathcal{U}_k^K(1)}{K} \Big) \Big].$$

By applying inequality (5.10) and Lemma 5.4 we succeed in bounding $\mathbb{P}^{(1)}(2R)$ by a constant over $\log K$. It remains to consider the event $R2a$ of a recombination within the a-population. Define the first time (with respect to the backwards in time process) that this event happens:

$$R_{aa}^{(1)}(i) := \sup\{m, m \leq J^K(1) \text{ and both neutral loci of the } i\text{-th individual are} \\ \text{associated to distinct } a\text{-individuals at the } (m-1)\text{th jump}, \tag{6.4}$$

where $R_{aa}^{(1)}(i) = -\infty$ if the event does not happen during the first phase of the sweep. Then,

$$\begin{aligned}
\mathbb{P}^{(1)}(R_{aa}^{(1)}(i) \geq 0) &= \sum_{l=1}^{\lfloor \varepsilon K \rfloor - 1} \mathbb{P}^{(1)}(R_{aa}^{(1)}(i) \geq 0, \tilde{N}_a(\tau^K_{R_{aa}^{(1)}(i)}) = l) \\
&= \sum_{l=1}^{\lfloor \varepsilon K \rfloor - 1} \sum_{m < \infty} \mathbb{P}^{(1)}(m \leq J^K(1), \tilde{N}_a(\tau^K_{m-1}) = l, \tilde{N}_a(\tau^K_m) = l+1, (ai1)_m, (ai2)_m, \forall m' > m : (ai12)_{m'}) \\
&\leq \sum_{l=1}^{\lfloor \varepsilon K \rfloor - 1} \sum_{m < \infty} \sup_{n_A \in I_\varepsilon^K \pm 1} \Big(p_{aa}^{(2)}(n_A, l) \mathbb{P}^{(1)}_{(n_A, l+1)}(\forall m \geq 0 : (ai12)_m) \Big) \mathbb{P}^{(1)}(m \leq J^K(1), \tilde{N}_a(\tau^K_{m-1}) = l, \tilde{N}_a(\tau^K_m) = l+1) \\
&\leq \sum_{l=1}^{\lfloor \varepsilon K \rfloor - 1} \frac{c}{K \log K} \mathbb{E}^{(1)}[U_l^K(1)] \leq \frac{c}{\log K},
\end{aligned}$$

by (5.10) and (6.2). □

To simplify the notations we will denote the union of all negligible events by

$$NE := aAa \cup CR \cup CA \cup 2R \cup R2a. \tag{6.5}$$

6.2. **The two loci of one individual separate within the A-population.** Having excluded events of small probability, there are exactly two ways for the neutral alleles of an individual sampled at the end of the first phase to originate from two distinct A-individuals. The two possibilities were already described on page 10 and represented in Figure 4. The ideas which are pursued in this section are similar to the ones from [2], but there are extra difficulties due to the randomness of the population size.

6.2.1. *Event* $[2,1]^{rec}_{A,i}$. The aim of this section is to prove the following approximation:

Proposition 5. *Let i be an a-individual sampled uniformly at the end of the first phase. There exist two finite constants c and ε_0 such that for $\varepsilon \leq \varepsilon_0$,*

$$\limsup_{K\to\infty} \Big|\mathbb{P}^{(1)}([2,1]^{rec}_{A,i}) - \Big[\frac{r_2}{r_1+r_2} - e^{-\frac{r_1}{s}\log\lfloor\varepsilon K\rfloor} + \frac{r_1}{r_1+r_2}e^{-\frac{r_1+r_2}{s}\log\lfloor\varepsilon K\rfloor}\Big]\Big| \leq c\sqrt{\varepsilon}.$$

We first give a preliminary Lemma before proving Proposition 5. Recall (4.2) and define for $j \in \{1,2,(1,2)\}$ and $m \in \mathbb{N}$,

$$\begin{aligned}R(i,j) := \sup\{m, m \leq J^K(1) &\text{ and the } j\text{-th locus/loci of the } i\text{-th individual}\\ &\text{is/are associated to an allele } A \text{ at the } (m-1)\text{th jump time}\},\end{aligned} \tag{6.6}$$

the last jump (forwards in time) when the j-th locus/loci of the i-th individual belongs to the A-population (with $\sup\emptyset = -\infty$). To prove Proposition 5 the idea is to decompose the event $[2,1]^{rec}_{A,i}$ according to the different possible a-population sizes when the first (backwards in time) recombination between $N1$ and $N2$ occurs.

$$\begin{aligned}\mathbb{P}^{(1)}([2,1]^{rec}_{A,i}) &= \mathbb{P}^{(1)}(R(i,2) > R(i,1) \geq 0)\\ &= \sum_{l=1}^{\lfloor\varepsilon K\rfloor} \mathbb{P}^{(1)}(R(i,1)\geq 0, R(i,2) > R(i,1), \tilde{N}_a(\tau^K_{R(i,2)}) = l)\\ &= \sum_{l=1}^{\lfloor\varepsilon K\rfloor-1} \mathbb{P}^{(1)}(R(i,2) > R(i,1), \tilde{N}_a(\tau^K_{R(i,2)}) = l)\mathbb{P}^{(1)}(R(i,1)\geq 0 | R(i,2) > R(i,1), \tilde{N}_a(\tau^K_{R(i,2)}) = l).\end{aligned} \tag{6.7}$$

In the following Lemma, which then gives rise to the proof of Proposition 5, we consider separately the two probabilities of the above product:

Lemma 6.4. *There exist three finite constants c, K_0 and ε_0 such that for $K \geq K_0$, $\varepsilon \leq \varepsilon_0$ and $l < \lfloor\varepsilon K\rfloor$,*

$$\Big|\mathbb{P}^{(1)}(R(i,2) > R(i,1), \tilde{N}_a(\tau^K_{R(i,2)}) = l) - r_2\frac{1-(1-s)^{\lfloor\varepsilon K\rfloor - l} - (1-s)^{l+1}}{s(l+1)}e^{-\frac{r_1+r_2}{s}\log\frac{\lfloor\varepsilon K\rfloor}{l}}\Big| \leq \frac{c\sqrt{\varepsilon}}{l\log K} \tag{6.8}$$

and

$$\Big|\mathbb{P}^{(1)}(R(i,1)\geq 0 | R(i,2) > R(i,1), N^K_a(\tau^K_{R(i,2)}) = l) - \sum_{k=1}^{l-1}\frac{r_1}{s(k+1)}e^{-\frac{r_1}{s}\log\frac{l-1}{k}}\Big| \leq c\sqrt{\varepsilon}. \tag{6.9}$$

Proof of Proposition 5. From Lemma 6.4 and Equation (6.7) we get the existence of a finite c such that for K large enough and ε small enough,

$$\mathbb{P}^{(1)}([2,1]^{rec}_{A,i}) \leq \sum_{l=1}^{\lfloor\varepsilon K\rfloor-1}\Big[\frac{r_2}{s(l+1)}e^{-\frac{r_1+r_2}{s}\log\frac{\lfloor\varepsilon K\rfloor}{l}} + \frac{c\sqrt{\varepsilon}}{l\log K}\Big]\Big[\sum_{k=1}^{l-1}\frac{r_1}{s(k+1)}e^{-\frac{r_1}{s}\log\frac{l-1}{k}} + c\sqrt{\varepsilon}\Big]. \tag{6.10}$$

Rewriting the second term in brackets and applying Lemma B.3 with $c_N/\log N = r_1/s$ yields:

$$\begin{aligned} e^{-\frac{r_1}{s}\log(l-1)}\frac{r_1}{s}\sum_{k=1}^{l-1}\frac{1}{k+1}e^{\frac{r_1}{s}\log k}+c\sqrt{\varepsilon} &\leq e^{-\frac{r_1}{s}\log(l-1)}\Big(e^{\frac{r_1}{s}\log l}-1+c\frac{r_1}{s}\Big)+c\sqrt{\varepsilon}\\ &\leq 1-e^{-\frac{r_1}{s}\log l}+c\sqrt{\varepsilon}, \end{aligned}$$

for K large enough, ε small enough and a finite c, whose value can change from line to line and which can be chosen independently of l. We use in the last inequality Condition (1.1) which claims that $\limsup_{K\to\infty} r_1 \log K < \infty$. Including the last inequality in (6.10) gives

$$\mathbb{P}^{(1)}([2,1]^{rec}_{A,i}) \leq \sum_{l=1}^{\lfloor\varepsilon K\rfloor-1}\frac{r_2}{s(l+1)}e^{-\frac{r_1+r_2}{s}\log\lfloor\varepsilon K\rfloor}\Big(e^{\frac{r_1+r_2}{s}\log l}-e^{\frac{r_2}{s}\log l}\Big)+c\sqrt{\varepsilon},$$

for a finite c, K large enough and ε small enough, where we again use (1.1) which ensures that exponential terms are bounded away from zero and infinity in the following sense:

$$\frac{1}{c}\leq\liminf_{K\to\infty}e^{-\frac{r_1+r_2}{s}\log\lfloor\varepsilon K\rfloor}\leq\limsup_{K\to\infty}e^{\frac{r_1+r_2}{s}\log\lfloor\varepsilon K\rfloor}\leq c$$

for a positive and finite c. Applying again Lemma B.3, we get:

$$\mathbb{P}^{(1)}([2,1]^{rec}_{A,i})\leq\Big(\frac{r_2}{r_1+r_2}-e^{-\frac{r_1}{s}\log\lfloor\varepsilon K\rfloor}+\frac{r_1}{r_1+r_2}e^{-\frac{r_1+r_2}{s}\log\lfloor\varepsilon K\rfloor}\Big)+c\sqrt{\varepsilon}.$$

The lower bound is obtained in the same way. Notice that it is a little bit more involved as we need to use (B.2) in addition. □

The end of this section is devoted to the proof of Lemma 6.4.

Proof of Equation (6.8). We can decompose the event $\{R(i,2)>R(i,1),\tilde{N}_a(\tau^K_{R(i,2)})=l\}$ according to the jump number of the (backwards in time) first recombination. Recall the definition of $NR(i)^{(1)}$ on page 10. We will use this event with a different initial condition for $(\tilde{N}_A,\tilde{N}_a)$, which will not necessarily be $(\lfloor\bar{n}_AK\rfloor,1)$. It will however still correspond to the absence of any recombination before the end of the first phase. We recall conventions (1.9) and (4.3). With the definition of $(\alpha ik)_m$ in (6.3) we get

$$\begin{aligned} &(6.11)\quad \mathbb{P}^{(1)}(R(i,2)>R(i,1),N_a(\tau^K_{R(i,2)})=l)\\ &=\sum_{m>1}\mathbb{P}^{(1)}(m\leq J^K(1),\tilde{N}_a(\tau^K_{m-1})=l-1,\tilde{N}_a(\tau^K_m)=l,(ai1)_{m-1},(Ai2)_{m-1},\forall m\leq m'\leq J^K(1):(ai12)_{m'})\\ &\leq\sum_{m>1}\sup_{n_A\in I^K_\varepsilon\pm1}\Big\{p^{(2)}_{aA}(n_A,l-1)\mathbb{P}^{(1)}_{(n_A,l)}(NR(i)^{(1)})\Big\}\mathbb{P}^{(1)}(m\leq J^K(1),\tilde{N}_a(\tau^K_{m-1})=l-1,\tilde{N}_a(\tau^K_m)=l)\\ &=\sup_{n_A\in I^K_\varepsilon\pm1}\Big\{p^{(2)}_{aA}(n_A,l-1)\mathbb{P}^{(1)}_{(n_A,l)}(NR(i)^{(1)})\Big\}\mathbb{E}^{(1)}[U^K_{l-1}(1)], \end{aligned}$$

and the same expression with the infimum on $n_A\in I^K_\varepsilon\pm1$ for a lower bound. Adding (6.2) and (A.1) yields,

$$\begin{aligned} \mathbb{P}^{(1)}(R(i,2)>R(i,1),\tilde{N}_a(\tau^K_{R(i,2)})=l) &\leq (1+c\varepsilon)\frac{r_2}{l+1}(e^{-\frac{r_1+r_2}{s}\log\frac{\lfloor\varepsilon K\rfloor}{l}}+c\sqrt{\varepsilon})\mathbb{E}^{(1)}[U^K_{l-1}(1)]\\ &\leq (1+c\sqrt{\varepsilon})\frac{r_2}{l+1}e^{-\frac{r_1+r_2}{s}\log\frac{\lfloor\varepsilon K\rfloor}{l}}\mathbb{E}^{(1)}[U^K_{l-1}(1)], \end{aligned}$$

for a finite c, ε small enough and K large enough, where we used that $(r_1+r_2)\log K$ is bounded. We similarly get a lower bound and end up the proof of Equation (6.8) by applying (5.10). □

Proof of Equation (6.9). We will decompose the event considered here according to the value of $\tilde{N}_a$ when the first (backwards in time) recombination occurs. Let us denote by $\zeta^K_k(1)$ the jump number of the last hitting of $k\leq\lfloor\varepsilon K\rfloor$ by $\tilde{N}_a$ during the first phase,

$$\zeta^K_k(1):=\sup\{m,\tau^K_m\leq\tilde{T}^K_\varepsilon,\tilde{N}_a(\tau^K_m)=k\},\tag{6.12}$$

and recall (5.19). Then we can define the events

$$NR(l,\xi,i) := \{\text{the first locus of individual } i \text{ sampled at jump time } \tau_\xi^K \text{ does not recombine from the } a\text{- to the } A\text{-population between 0 and } \tau_\xi^K\} \tag{6.13}$$

where $\xi \in \{\zeta_l^K(1), \sigma_l^K(1)\}$. Similarly as in (6.11), Bayes' rule leads to:

$$\begin{aligned}
&\mathbb{P}^{(1)}(R(i,1)\geq 0 \mid R(i,2)>R(i,1), \tilde{N}_a(\tau^K_{R(i,2)})=l) \\
&= \sum_{k=1}^{\lfloor \varepsilon K\rfloor} \mathbb{P}^{(1)}(R(i,1)\geq 0, \tilde{N}_a(\tau^K_{R(i,1)})= k \mid R(i,2)>R(i,1), \tilde{N}_a(\tau^K_{R(i,2)})=l), \\
&\leq \sum_{k=1}^{\lfloor \varepsilon K\rfloor} \Big(\sup_{n_A\in I_\varepsilon^K\pm 1} p_{aA}^{(1)}(n_A,k-1)\mathbb{P}^{(1)}_{(n_A,k)}(NR(l,\sigma,i))\Big)\mathcal{S}(k,l),
\end{aligned} \tag{6.14}$$

where for the sake of simplicity we have introduced the notation

$$\mathcal{S}(k,l) := \sum_{m<\infty} \mathbb{P}^{(1)}(m<R(i,2), \tilde{N}_a(\tau^K_{m-1})=k-1, \tilde{N}_a(\tau^K_m)=k \mid \tilde{N}_a(\tau^K_{R(i,2)})=l).$$

The lower bound is obtained by taking the infimum for n_A in $I_\varepsilon^K\pm 1$ and replacing σ by ζ. To lighten the proof, we bound the probability in the brackets for both σ and ζ in Lemma A.1, Equation (A.2).

First we prove that with a probability close to one the a-population size is bigger when the (backwards in time) first recombination occurs than when the second, of locus $(i,1)$, occurs. Note that by (5.10) and Lemma 5.3, there exists a finite c such that for every $l<k<\lfloor\varepsilon K\rfloor$:

$$\mathcal{S}(k,l) \leq \mathbb{E}^{(1)}[U_l^K(1)] \sup_{n_A\in I_\varepsilon^K\pm 1} \mathbb{E}^{(1)}_{(n_A,l+1)}[U^K_{n_A,l,k-1}(1)|\sigma_l^K(1)<\infty] \leq c\mu_\varepsilon^{k-l},$$

where we recall that $\mu_\varepsilon<1$ for ε small enough. Hence, recalling (6.14) and (6.2), we obtain for $k>l$

$$\mathbb{P}^{(1)}(R(i,1)\geq 0, \tilde{N}_a^K(\tau^K_{R(i,1)})\geq l | R(i,2)>R(i,1), \tilde{N}_a(\tau^K_{R(i,2)})=l) \leq cr_1 \sum_{k=l+1}^{\lfloor\varepsilon K\rfloor} \frac{\mu_\varepsilon^{k-l}}{k} \leq \frac{c}{\log K},$$

for a finite c and ε small enough, which entails

$$\begin{aligned}
&\mathbb{P}^{(1)}(R(i,1)\geq 0 \mid R(i,2)>R(i,1), \tilde{N}_a(\tau^K_{R(i,2)})=l) \\
&\qquad \leq \sum_{k=1}^{l} \Big(\sup_{n_A\in I_\varepsilon^K\pm 1} p_{aA}^{(1)}(n_A,k-1)\mathbb{P}^{(1)}_{(n_A,k)}(NR(l,\sigma,i))\Big)\mathcal{S}(k,l) + O\left(\frac{1}{\log K}\right).
\end{aligned}$$

We therefore can ignore all $k>l$ in the sum in (6.14) and continue with the case $k\leq l$. In this setting, we can bound the sum $\mathcal{S}(k,l)$ as follows:

$$\mathbb{E}^{(1)}[U^K_{k-1}(1)] - \sup_{n_A\in I_\varepsilon^K\pm 1} \mathbb{E}^{(1)}_{(n_A,l-1)}[U^K_{n_A,l,k-1}(1)]\mathbb{E}^{(1)}[U_l^K(1)] \leq \mathcal{S}(k,l) \leq \mathbb{E}^{(1)}[U^K_{k-1}(1)].$$

Bounding the difference between the two bounds above within Equation (6.14) then yields

$$\sum_{k=1}^{l} \frac{r_1}{k} \sup_{n_A\in I_\varepsilon^K\pm 1} \mathbb{E}^{(1)}_{(n_A,l-1)}[U^K_{n_A,l,k-1}(1)]\mathbb{E}^{(1)}[U_l^K(1)] \leq cr_1\sum_{k=1}^{l}\frac{\mu_\varepsilon^{l-k}}{k} \leq \frac{c}{\log K},$$

for a finite c by (6.2), (5.10) and Lemma 5.3. As a consequence,

$$\sum_{k=1}^{l} \Big(\sup_{n_A\in I_\varepsilon^K\pm 1} p_{aA}^{(1)}(n_A,k-1)\mathbb{P}^{(1)}_{(n_A,k)}(NR(l,\sigma,i))\Big)\Big|\mathcal{S}(k,l) - \mathbb{E}^{(1)}[U^K_{k-1}(1)]\Big| \leq O\left(\frac{1}{\log K}\right),$$

and thus we can work with $\mathbb{E}^{(1)}[U^K_{k-1}(1)]$ as an approximation for the sum $\mathcal{S}(k,l)$:

$$\begin{aligned}
&\mathbb{P}^{(1)}(R(i,1)\geq 0 \mid R(i,2)>R(i,1), \tilde{N}_a(\tau^K_{R(i,2)})=l) \\
&\qquad \leq \sum_{k=1}^{l} \Big(\sup_{n_A\in I_\varepsilon^K\pm 1} p_{aA}^{(1)}(n_A,k-1)\mathbb{P}^{(1)}_{(n_A,k)}(NR(l,\sigma,i))\Big)\mathbb{E}^{(1)}[U^K_{k-1}(1)] + O\left(\frac{1}{\log K}\right).
\end{aligned}$$

Reasoning in the same way to get a lower bound and using (6.2) and (A.2) we get the existence of a finite c such that for K large enough and ε small enough,

$$\Big|\mathbb{P}^{(1)}(R(i,1)\geq 0|R(i,2)>R(i,1),\tilde{N}_a^K(\tau^K_{R(i,2)})=l)-\sum_{k=1}^{l-1}\frac{r_1}{k}e^{-\frac{r_1}{s}\log\frac{l-1}{k}}\mathbb{E}^{(1)}[U_k^K(1)]\Big|\leq c\sqrt{\varepsilon}.$$

Applying (5.10) and (B.2) yields Equation (6.9). Notice that we have replaced $1/k$ by $1/(k+1)$. We used Condition (1.1) to do this. $\square$

6.2.2. *Event* $[12,2]^{rec}_{A,i}$. Recall the definition of $[12,2]^{rec}_{A,i}$ on page 10. This section is devoted to the proof of the following result:

Proposition 6. *Let i be an individual sampled uniformly at the end of the first phase. There exist two finite constants c and ε_0 such that for $\varepsilon\leq\varepsilon_0$,*

$$\limsup_{K\to\infty}\Big|\mathbb{P}^{(1)}([12,2]^{rec}_{A,i})-r_1\Big[\frac{1-e^{-\frac{r_1+r_2}{s}\log\lfloor\varepsilon K\rfloor}}{r_1+r_2}+\frac{e^{-\frac{r_1+r_2}{s}\log\lfloor\varepsilon K\rfloor}-e^{-\frac{f_Ar_2}{f_as}\log\lfloor\varepsilon K\rfloor}}{r_1+r_2(1-f_A/f_a)}\Big]\Big|\leq c\sqrt{\varepsilon}.$$

Proof. As the proof is very similar to the proof of Proposition 5 we will be very brief here and only give the ingredients. Let us introduce for $l<\lfloor\varepsilon K\rfloor$ the event:

$$RA(l,i):=\{[12,2]^{rec}_{A,i}\mid R(i,2)=R(i,1)\geq 0,\tilde{N}_a(\tau^K_{R(i,1)})=l\}. \tag{6.15}$$

Then we can rewrite the probability of $[12,2]^{rec}_{A,i}$ as follows:

$$\mathbb{P}^{(1)}([12,2]^{rec}_{A,i})=\sum_{l=1}^{\lfloor\varepsilon K\rfloor}\mathbb{P}^{(1)}(RA(l,i))\mathbb{P}^{(1)}(R(i,2)=R(i,1)\geq 0,\tilde{N}_a^K(\tau^K_{R(i,1)})=l). \tag{6.16}$$

Apart from the point of recombination, the second probability in the above sum coincides with the probability studied in (6.8) and we obtain for ε small enough and K large enough,

$$\sup_{l\leq\lfloor\varepsilon K\rfloor} l\cdot\Big|\mathbb{P}^{(1)}(R(i,2)=R(i,1)\geq 0,N_a^K(\tau^K_{R(i,1)})=l)-\frac{r_1(1-(1-s)^{\lfloor\varepsilon K\rfloor-l}-(1-s)^{l+1})}{s(l+1)}e^{-\frac{r_1+r_2}{s}\log\frac{\lfloor\varepsilon K\rfloor}{l}}\Big|\leq c\frac{\sqrt{\varepsilon}}{\log K}, \tag{6.17}$$

for a finite c, when substituting r_2 by r_1 in the fraction which mirrors the recombination probability. The probability of $RA(l,i)$ is derived in Lemma A.1. Inserting (6.17) and (A.3) into (6.16) yields

$$\begin{aligned}\mathbb{P}^{(1)}([12,2]^{rec}_{A,i}) &\leq \sum_{l=1}^{\lfloor\varepsilon K\rfloor}(1-e^{-\frac{f_A}{f_a}\frac{r_2}{s}\log l})\frac{r_1}{l+1}e^{-\frac{r_1+r_2}{s}\log\frac{\lfloor\varepsilon K\rfloor}{l}}+c\sqrt{\varepsilon}\\ &\leq r_1e^{-\frac{r_1+r_2}{s}\log\lfloor\varepsilon K\rfloor}\Big[\frac{e^{\frac{r_1+r_2}{s}\log\lfloor\varepsilon K\rfloor}-1}{r_1+r_2}-\frac{e^{\frac{r_1+r_2-f_Ar_2/f_a}{s}\log\lfloor\varepsilon K\rfloor}-1}{r_1+r_2-f_Ar_2/f_a}\Big]+c\sqrt{\varepsilon}\end{aligned}$$

where we again applied Lemma B.3 to express the sum in a different way, and used the finiteness of $\limsup_{K\to\infty}(r_1+r_2)\log K$ assumed in Condition (1.1). Reasoning similarly for the lower bound and rearranging the terms end the proof of Proposition 6. $\square$

6.3. **Proof of Proposition 1.**
Event $R2(i)^{(1)}$: By definition and from Lemma 6.3,

$$\mathbb{P}^{(1)}(R2(i)^{(1)})=\mathbb{P}^{(1)}(R(i,2)\geq 0)-\mathbb{P}^{(1)}(R(i,1)\geq 0)+O\Big(\frac{\log K}{K}\Big),$$

where $R(i,1)$ and $R(i,2)$ have been defined in (6.6). But these probabilities have already been derived in [19] Lemma 7.4, and we get:

$$\mathbb{P}^{(1)}(R2(i)^{(1)})=(1-q_1q_2)-(1-q_1)+O_K(\varepsilon)=q_1(1-q_2)+O_K(\varepsilon),$$

where $O_K(\varepsilon)$ satisfies (4.7).
Event $R1|2(i)^{(1,ga)}$: By definition (see page 10)

$$\mathbb{P}^{(1)}(R1|2(i)^{(1,ga)})=\mathbb{P}^{(1)}([2,1]^{rec}_{A,i})+\mathbb{P}^{(1)}([12,2]^{rec}_{A,i}).$$

The result then follows from Propositions 5 and 6.

Event $R12(i)^{(1)}$: From Definition (6.15) and Equation (A.3) we obtain for K large enough,

$$\begin{aligned}
\mathbb{P}^{(1)}(R12(i)^{(1)}) &= \sum_{l=1}^{\lfloor \varepsilon K \rfloor}(1-\mathbb{P}^{(1)}(RA(l,i)))\mathbb{P}^{(1)}(R(i,1)=R(i,2)\geq 0, \tilde{N}_a(\tau^K_{R(i,2)})=l)\\
&= r_1 \sum_{l=1}^{\lfloor \varepsilon K \rfloor} e^{-\frac{f_A}{f_a}\frac{r_2}{s}\log l}\frac{1-(1-s)^{\lfloor \varepsilon K\rfloor - l}-(1-s)^{l+1}}{s(l+1)} e^{-\frac{r_1+r_2}{s}\log\frac{\lfloor \varepsilon K\rfloor}{l}} + O_K(\sqrt{\varepsilon})\\
&= \frac{r_1}{r_1+r_2-f_A r_2/f_a}\Big(e^{-\frac{r_2}{s}\frac{f_A}{f_a}\log\lfloor \varepsilon K\rfloor} - e^{-\frac{r_1+r_2}{s}\log\lfloor \varepsilon K\rfloor}\Big) + O_K(\sqrt{\varepsilon}),
\end{aligned}$$

where we again used the statement of Lemma B.3 to substitute the sum, as well as Equation (B.2).

Event $NR(i)^{(1)}$: From Lemma 6.3,

$$\mathbb{P}^{(1)}(NR(i)^{(1)}) = 1-\mathbb{P}^{(1)}(R2(i)^{(1)}) - \mathbb{P}^{(1)}(R12(i)^{(1)}) - \mathbb{P}^{(1)}(R2(i)^{(1)}) + O\Big(\frac{\log K}{K}\Big).$$

This ends up the proof of Proposition 1. □

7. Second and third phases

This section is devoted to the proofs of Propositions 2 and 3.

7.1. **Proof of Proposition 2.** We need to show that two distinct lineages picked uniformly at the end of the second phase coalesce or recombine during that phase only with negligible probability. Let us recall the definition of the jumps τ^K_m in (4.1) and denote by $U^K(2)$ the number of upcrossings of the a-population during the second phase:

$$U^K(2) := \#\{m, T^K_\varepsilon < \tau^K_m \leq T^K_\varepsilon + t_\varepsilon, N_a(\tau^K_{m+1}) - N_a(\tau^K_m) = 1\}. \tag{7.1}$$

Let us introduce the event C^K_ε:

$$C^K_\varepsilon := \{T^K_\varepsilon \leq S^K_\varepsilon\} \cap \{N^K_a(t) \geq \varepsilon^2 K/4, \forall\, T^K_\varepsilon \leq t \leq T^K_\varepsilon + t_\varepsilon\}.$$

In particular on the event C^K_ε, for $T^K_\varepsilon \leq \tau^K_m \leq T^K_\varepsilon + t_\varepsilon$ and $j \in \{1,2\}$

$$p^{(j)}_{aA}(N(\tau^K_m)) \leq \frac{8r_j}{\varepsilon^2 K} \quad \text{and} \quad p^{(c,j)}_{aa}(N(\tau^K_m)) \leq \frac{32}{\varepsilon^4 K^2}.$$

Then if we recall the definition of $NR(i)^{(2)}$ on page 11 we have for $m \in \mathbb{N}$,

$$\mathbb{P}^{(1)}(NR(i)^{(2)} | U^K(2) = m, C^K_\varepsilon) \geq \Big(1 - \frac{8(r_1+r_2)}{\varepsilon^2 K}\Big)^m. \tag{7.2}$$

But for K large enough, $\log(1-8(r_1+r_2)/(\varepsilon^2 K)) \geq -10(r_1+r_2)/(\varepsilon^2 K)$ and hence

$$\begin{aligned}
\mathbb{P}^{(1)}(NR(i)^{(2)}|C^K_\varepsilon) &\geq \Big(1-\mathbb{P}^{(1)}(U^K(2) > K\log\log K | C^K_\varepsilon)\Big) e^{K\log\log K \log(1-\frac{8(r_1+r_2)}{\varepsilon^2 K})}\\
&\geq \Big(1-\mathbb{P}^{(1)}(U^K(2) > K\log\log K | C^K_\varepsilon)\Big) e^{-\frac{10(r_1+r_2)\log\log K}{\varepsilon^2}}.
\end{aligned}$$

According to Condition (1.1) the exponential term is equivalent to 1 when K is large. Moreover, by (3.5), N^K_a is smaller than $2\bar{n}_a K$ on the time interval $[T^K_\varepsilon, T^K_\varepsilon + t_\varepsilon]$ with probability close to 1. When this property holds, we can bound the birth number $U^K(2)$ by the sum of $2\bar{n}_a K$ iid Poisson random variables with parameter $f_a t_\varepsilon$. The strong law of large numbers then yields

$$\lim_{K\to\infty} \mathbb{P}^{(1)}(U^K(2) > K\log\log K | C^K_\varepsilon) = 0.$$

Applying again (3.5) to get $\lim_{K\to\infty}\mathbb{P}(C^K_\varepsilon | T^K_\varepsilon < \infty) = 1$ finally gives

$$\lim_{K\to\infty} \mathbb{P}(NR(i)^{(2)} | T^K_\varepsilon < \infty) = 1.$$

The coalescence part in Proposition 2 can be proven in the same way.

7.2. **Proof of Proposition 3.** The proof of the asymptotic probability of $R2(i)^{(3,ga)}$ is the same as for (A.3), except that the roles of A and a are exchanged. Hence we do not give more details. Note however that it extensively uses Lemma 5.6. Let us now focus on the event $NR(i)^{(3)}$, and introduce

$$NRA(i)^{(3)} := \{\text{no neutral allele of individual } i \text{ recombines from the } a \text{ to the } A \text{ population}\}.$$

Recall the definitions of $\mathbb{P}^{(3)}$ and $U^K(3)$ in (4.3) and (5.38) respectively. We decompose the probabilities according to the number of upcrossings of $\tilde{\tilde{N}}_a$ during the third phase and get in the same way as in (7.2), for $m \in \mathbb{N}$

$$\mathbb{P}^{(3)}(NRA(i)^{(3)}|U^K(3)=m, \{\tilde{\tilde{T}}_0^{(K,A)} < \tilde{\tilde{T}}_\varepsilon^{(K,A)} \wedge S_\varepsilon^{(K,a)}\}) \geq \Big(1-\frac{f_A(r_1+r_2)\varepsilon}{f_a(\bar{n}_a - M''\varepsilon)^2 K}\Big)^m,$$

where we recall that $\tilde{\tilde{T}}_0^{(K,A)}$ and $\tilde{\tilde{T}}_\varepsilon^{(K,A)}$ are the analogs of $T_0^{(K,A)}$ and $T_\varepsilon^{(K,A)}$ (defined in (3.11)) for the process $\tilde{\tilde{N}}$. But for K large enough and ε small enough,

$$\log\Big(1-\frac{f_A(r_1+r_2)\varepsilon}{f_a(\bar{n}_a-M''\varepsilon)^2K}\Big) \geq -2f_A\frac{(r_1+r_2)\varepsilon}{f_a\bar{n}_a^2K}.$$

Hence we get for a finite constant c and ε small enough:

$$\begin{aligned}\mathbb{P}^{(3)}(NRA(i)^{(3)}) \quad &\geq \quad \Big(1-\mathbb{P}^{(3)}\Big(U^K(3)>\frac{K\log K}{\sqrt{\varepsilon}}\Big)\Big)\exp\Big(-\frac{2f_A(r_1+r_2)\sqrt{\varepsilon}\log K}{f_a\bar{n}_a^2}\Big)\\ &\geq \quad \Big(1-\frac{\sqrt{\varepsilon}\mathbb{E}^{(3)}[U^K(3)]}{K\log K}\Big)\Big(1-\frac{2f_A(r_1+r_2)\sqrt{\varepsilon}\log K}{f_a\bar{n}_a^2}\Big) \geq (1-c\sqrt{\varepsilon})^2,\end{aligned}$$

where we used Lemma 5.6 and that $(r_1+r_2)\log K$ is bounded (Condition (1.1)).

The proof of the last part of Proposition 3 is very similar to that of Proposition 2. The key arguments are that the expectation of the birth number of a-individuals during the third phase under $\mathbb{P}^{(3)}$ is of order $K\log K$ (Lemma 5.6), whereas the probability for two neutral alleles associated with an allele a to coalesce is of order $1/K^2$ at each birth of an a-individual (Lemma 6.1).

8. Independence of neutral lineages

This section is dedicated to the proof of Proposition 4. We sample d distinct individuals uniformly at the end of the first phase. We recall the definitions of the genealogical events during the first phase on page 10 and introduce:

$$R(1|2) := \sum_{1\leq i\leq d}\mathbf{1}_{R1|2(i)^{(1,ga)}}, \quad R(1) := R(1|2) + \sum_{1\leq i\leq d}\mathbf{1}_{R12(i)^{(1)}} \quad \text{and} \quad R(2) := R(1) + \sum_{1\leq i\leq d}\mathbf{1}_{R2(i)^{(1)}}$$

From Proposition 1 we know that $R(1)$, $R(2)$ and $R(1|2)$ are sufficient to describe the neutral genealogies at the end of the first phase up to a probability negligible with respect to one for large K. Let j,k,l be three integers such that $l \leq j$ and $j+k \leq d$. We aim at approximating

$$\begin{aligned}p(j,k,l) :&= \mathbb{P}(R(1)=j, R(2)=j+k, R(1|2)=l|T_\varepsilon^K \leq S_\varepsilon^K)\\ &= \mathbb{P}(R(1)=j|T_\varepsilon^K\leq S_\varepsilon^K)\mathbb{P}(R(2)=j+k|T_\varepsilon^K\leq S_\varepsilon^K, R(1)=j)\\ &\quad\ \mathbb{P}(R(1|2)=l|T_\varepsilon^K\leq S_\varepsilon^K, R(1)=j, R(2)=j+k).\end{aligned} \tag{8.1}$$

The approximations of the two first probabilities are direct adaptations of Lemma 5.2 and the proof of Proposition 2.6 in [18], pp 1623-1624. More precisely, Lemma 7.3 in [19] which states that with high probability two neutral lineages do not coalesce and then recombine (backwards in time) allows us to get an equivalent of Lemma 5.2 (with $J=0$) in [18]:

$$\Big|\mathbb{P}(R(1)=j|T_\varepsilon^K\leq S_\varepsilon^K) - \binom{d}{j}\mathbb{E}[F_1^j(1-F_1)^{n-j}|T_\varepsilon^K\leq S_\varepsilon^K]\Big| \leq c\Big(\frac{1}{\log K}+\varepsilon\Big),$$

for ε small enough, where c is a finite constant,

$$F_1 := \mathbb{P}(R(i,1)\geq 0|((N_A,N_a)(\tau_n^K), n\leq J^K(1)), T_\varepsilon^K\leq S_\varepsilon^K),$$

and $R(i,1)$ is defined in (6.6). Then Equations (7.21), (7.23), (7.24) and (7.26) of [19] yield

$$\limsup_{K\to\infty} \Big|\mathbb{E}[F_1^j(1-F_1)^{d-j}|T_\varepsilon^K \leq S_\varepsilon^K] - (1-q_1)^j q_1^{(n-j)}\Big| \leq c\varepsilon,$$

where q_1 has been defined in (1.12), which allows to conclude

$$\limsup_{K\to\infty} \Big|\mathbb{P}(R(1)=j|T_\varepsilon^K \leq S_\varepsilon^K) - \binom{d}{j}(1-q_1)^j q_1^{(d-j)}\Big| \leq c\varepsilon, \tag{8.2}$$

for ε small enough where c is a finite constant.

The derivation of the second probability, $\mathbb{P}(R(2)=j+k|T_\varepsilon^K \leq S_\varepsilon^K, R(1)=j)$, follows the same outline. The lineages where $N1$ does not escape the sweep can be seen as lineages where SL and $N1$ are the same locus and the recombination probability between $SL-N1$ and $N2$ is r_2. This is due to the independence of the recombinations between SL and $N1$ and between $N1$ and $N2$. Hence we can rewrite the probability as follows:

$$\mathbb{P}(R(2)=j+k|T_\varepsilon^K \leq S_\varepsilon^K, R(1)=j) = \mathbb{P}(R(2)-R(1)=k|T_\varepsilon^K \leq S_\varepsilon^K, d-R(1)=d-j).$$

We can then directly apply the result (8.2) for the law of $R(1)$ and get:

$$\limsup_{K\to\infty} \Big|\mathbb{P}(R(2)=j+k|T_\varepsilon^K \leq S_\varepsilon^K, R(1)=j) - \binom{d-j}{k}(1-q_2)^k q_2^{(d-j-k)}\Big| \leq c\varepsilon, \tag{8.3}$$

for ε small enough where c is a finite constant and q_2 has been defined in (1.12).

The derivation of the last probability in (8.1) is more involved but follows the same spirit. First note that we only have to focus on genealogies where $N1$ escapes the sweep. Hence the derivation of the probability comes down to the derivation of $\mathbb{P}(R(1|2)=l|T_\varepsilon^K \leq S_\varepsilon^K, R(1)=j)$. The idea is to propose an alternative construction of the process with the same law and where we add the recombinations between $N1$ and $N2$ at the end:

- First we construct a trait population process (N_A, N_a) with birth and death rates defined in (1.4)
- Second we "add" the recombinations between SL and $N1$: at each birth event we draw a Bernoulli variable with parameter r_1 to decide whether there is a recombination or not. If there is a recombination, the parent giving its neutral allele at $N1$ is chosen with a probability proportional to its fertility (f_A or f_a).
  After this step of the construction we know the genealogies of the d neutral alleles at $N1$ sampled at the end of the sweep. We label $(i_1, ..., i_j)$ the j sampled neutral alleles at $N1$ which experience a recombination between SL and $N1$ in their genealogy.
- Third we "add" the recombinations between $N1$ and $N2$ sequentially in the lineages where there is already a recombination between SL and $N1$: we first follow backward in time the lineage of i_1 and at each birth event we draw a Bernoulli variable with parameter r_2 to decide whether there is a recombination or not, and choose the parent of neutral allele at $N2$ as in the second step. Then we do the same with the lineage of i_2, and so on until the lineage of i_j.
- Finally we "add" the recombinations between $N1$ and $N2$ in those lineages which were not marked with any recombination between SL and $N1$.

Such a construction generates a process distributed as the original process and facilitates the study of the dependencies between lineages $(i_1, ..., i_j)$. According to Lemma 6.3, with high probability there is no recombination between SL and $N1$ after (backwards in time) a coalescence at locus $N1$ among the d sampled individuals. In the same way, there is no coalescence at locus $N1$ after a recombination between SL and $N1$ in the A-population (this is due to the large number of A-individuals; similar proof as for the last probability of Proposition 3. Hence if we introduce

$$NC(j) := \{\text{there is no coalescence between lineages } (i_1, ..., i_j) \text{ at locus } N1\},$$

we get:

$$\mathbb{P}(R(1|2)=l|T_\varepsilon^K \leq S_\varepsilon^K, R(1)=j) = \mathbb{P}(R(1|2)=l|T_\varepsilon^K \leq S_\varepsilon^K, R(1)=j, NC(j)) + O\Big(\frac{\log K}{K}\Big).$$

With the construction of the alternative process we can also define sequentially for $1 \leq k \leq j$:

$$NC(j,k) := \{\text{there is no coalescence between lineages } (i_1,...,i_k) \text{ after completion of the process of adding the recombinations between } N1 \text{ and } N2 \text{ in the lineage } i_k\}.$$

Then, if we introduce for $1 \leq k \leq j$ and $\delta \in \{0,1\}$

$$\{r_{i_k} = \delta\} := \{\text{there is } \delta \text{ recombination between } N1 \text{ and } N2 \text{ in the lineage } i_k\},$$

then for $(\delta_1,...,\delta_j) \in \{0,1\}^j$

$$\mathbb{P}(r_{i_k} = \delta_k, 1 \leq k \leq j | T^K_\varepsilon \leq S^K_\varepsilon, R(1) = j) = \prod_{1\leq k\leq j} \mathbb{P}(r_{i_k} = \delta_k | T^K_\varepsilon \leq S^K_\varepsilon, R(1) = j, NC(j), NC(j,1),...,NC(j,k-1)) + O\Big(\frac{\log K}{K}\Big).$$

Indeed, the probability that the event $NC(j,k)$ is not realized after witnessing the recombinations between $N1$ and $N2$ in lineage i_k has order $\log K/K$ according to Lemma 6.3. But for $1 \leq k \leq j$,

$$\begin{aligned}
(8.4)\quad & \mathbb{P}(r_{i_k} = \delta_k | T^K_\varepsilon \leq S^K_\varepsilon, R(1) = j, NC(j),...,NC(j,k-1)) \\
&= \frac{\mathbb{P}(r_{i_k} = \delta_k, R(1) = j, NC(j),...,NC(j,k-1) | T^K_\varepsilon \leq S^K_\varepsilon)}{\mathbb{P}(R(1) = j, NC(j),...,NC(j,k-1) | T^K_\varepsilon \leq S^K_\varepsilon)} \\
&= \frac{\mathbb{P}(r_{i_k} = \delta_k, R(1) = j | T^K_\varepsilon \leq S^K_\varepsilon) - \mathbb{P}(r_{i_k} = \delta_k, R(1) = j, (NC(j)\cap...\cap NC(j,k-1))^c | T^K_\varepsilon \leq S^K_\varepsilon)}{\mathbb{P}(R(1) = j | T^K_\varepsilon \leq S^K_\varepsilon) - \mathbb{P}(R(1) = j, (NC(j)\cap...\cap NC(j,k-1))^c | T^K_\varepsilon \leq S^K_\varepsilon)},
\end{aligned}$$

and according to Lemma 6.3 and Coupling (3.14), there exists a finite c such that for K large enough and ε small enough,

$$\mathbb{P}((NC(j)\cap...\cap NC(j,k-1))^c | T^K_\varepsilon \leq S^K_\varepsilon) \leq c\Big(\frac{\log K}{K} + \varepsilon\Big).$$

As $\mathbb{P}(r_{i_k} = \delta_k, R(1) = j | T^K_\varepsilon \leq S^K_\varepsilon)$ does not go to 0 when K goes to infinity, we get

$$\begin{aligned}
&\mathbb{P}(r_{i_k} = \delta_k | T^K_\varepsilon \leq S^K_\varepsilon, R(1) = j, NC(j),...,NC(j,k-1)) = \mathbb{P}(r_{i_k} = \delta_k | T^K_\varepsilon \leq S^K_\varepsilon, R(1) = j) + O\Big(\frac{\log K}{K} + \varepsilon\Big) \\
&= \delta_k \frac{\mathbb{P}(R1|2(i_k)^{(1,ga)} | T^K_\varepsilon \leq S^K_\varepsilon)}{\mathbb{P}(R(i_k,1) \geq 0 | T^K_\varepsilon \leq S^K_\varepsilon)} + (1-\delta_k)\Big(1 - \frac{\mathbb{P}(R1|2(i_k)^{(1,ga)} | T^K_\varepsilon \leq S^K_\varepsilon)}{\mathbb{P}(R(i_k,1) \geq 0 | T^K_\varepsilon \leq S^K_\varepsilon)}\Big) + O\Big(\frac{\log K}{K} + \varepsilon\Big) \\
&= \delta_k \frac{1-q_1-q_3}{1-q_1} + (1-\delta_k)\Big(1 - \frac{1-q_1-q_3}{1-q_1}\Big) + O\Big(\frac{\log K}{K} + \varepsilon\Big),
\end{aligned}$$

where we recall the definition of $R(i_k,1)$ in (6.6), the definition of $R1|2(i_k)^{(1,ga)}$ on page 10, and we used Proposition 1. Adding Equations (8.2) and (8.3) we finally obtain:

$$\begin{aligned}
p(j,k,l) &= \binom{n}{j}(1-q_1)^j q_1^{(n-j)} \binom{n-j}{k} (1-q_2)^k q_2^{(n-j-k)} \binom{j}{l} \Big(1 - \frac{q_3}{1-q_1}\Big)^l \Big(\frac{q_3}{1-q_1}\Big)^{j-l} + O_K(\varepsilon) \\
(8.5)\quad &= \frac{n!}{l!(j-l)!k!(n-j-k)!}(q_1q_2)^{n-j-k}(q_1(1-q_2))^k q_3^{j-l}(1-q_1-q_3)^l + O_K(\varepsilon).
\end{aligned}$$

This ends the proof of the independence between genealogies during the first phase.

The derivation of the asymptotic independence of neutral lineages during the third phase is an easy adaptation of Lemma 5.2 and the proof of Proposition 2.6 in [18], pp 1623-1624 as with high probability two lineages do not coalesce during this phase. □

Appendix A. Lemma A.1

Recall the definition of $NR(i)^{(1)}$ on page 10, and Definitions (6.13) and (6.15). Then we have the following approximations for large K.

Lemma A.1. *There exist three finite constants c, K_0 and ε_0 such that for every $K \geq K_0$ and $\varepsilon \leq \varepsilon_0$*

$$(A.1)\quad \sup_{n_A \in I^K_\varepsilon \pm 1, l \leq \lfloor \varepsilon K \rfloor} \Big| \mathbb{P}^{(1)}_{(n_A,l)}(NR(i)^{(1)}) - \exp\Big(-\frac{r_1+r_2}{s}\log\frac{\lfloor \varepsilon K\rfloor}{l}\Big)\Big| \leq c\sqrt{\varepsilon},$$

$$\text{(A.2)} \qquad \sup_{\tau\in\{\zeta,\sigma\}} \sup_{n_A\in I_\varepsilon^K\pm 1, k\le l\le \lfloor \varepsilon K\rfloor} \Big|\mathbb{P}^{(1)}_{(n_A,k)}(NR(l,\tau,i) - \exp\Big(-\frac{r_1}{s}\log\frac{l-1}{k}\Big)\Big| \le c\sqrt{\varepsilon},$$

$$\text{(A.3)} \qquad \sup_{l\le\lfloor \varepsilon K\rfloor} \Big|\mathbb{P}^{(1)}(RA(l,i)) - \Big(1-\exp\Big(-\frac{f_A}{f_a}\frac{r_2}{s}\log l\Big)\Big)\Big| \le c\sqrt{\varepsilon}.$$

Proof. Let us introduce the sigma-algebra generated by the trait population process

$$\mathcal{F} := \sigma\Big((\tilde{N}_A,\tilde{N}_a)(\tau_m^K), \tau_m^K \le \tilde{T}_\varepsilon^K\Big).$$

We use some ideas developed in [18] and extended in [2] towards the two-locus case. The proof, although quite technical, can be summarized easily: for $(g,b,c,d,f)\in\mathbb{R}_+^5$, the Triangle Inequality and the Mean Value Theorem imply

$$|g-e^{-b}| \le |g-e^{-c}| + |c-d| + |d-f| + |f-b|.$$

Hence for every random variables $(X_1,X_2)\in\mathbb{R}_+^2$ and measurable event C:

$$\Big|\mathbb{P}^{(1)}(C|\mathcal{F}) - e^{-\frac{r_1+r_2}{s}\log\frac{\lfloor\varepsilon K\rfloor}{l}}\Big| \le \Big|\mathbb{P}^{(1)}(C|\mathcal{F}) - e^{-X_1}\Big| + \Big|X_1-X_2\Big| + \Big|X_2-\mathbb{E}^{(1)}[X_2]\Big| + \Big|\mathbb{E}^{(1)}[X_2] - \frac{r_1+r_2}{s}\log\frac{\lfloor\varepsilon K\rfloor}{l}\Big|.$$

By taking the expectation and applying Jensen and Cauchy-Schwarz Inequalities, we obtain:

$$\text{(A.4)} \quad \Big|\mathbb{P}^{(1)}(C) - e^{-\frac{r_1+r_2}{s}\log\frac{\lfloor\varepsilon K\rfloor}{l}}\Big| \le \mathbb{E}^{(1)}\Big|\mathbb{P}^{(1)}(C|\mathcal{F}) - e^{-X_1}\Big| + \mathbb{E}^{(1)}\Big|X_1-X_2\Big| + \sqrt{\mathrm{Var}(X_2)} + \Big|\mathbb{E}^{(1)}[X_2] - \frac{r_1+r_2}{s}\log\frac{\lfloor\varepsilon K\rfloor}{l}\Big|.$$

Hence the idea is to find the appropriate random variables $(X_1,X_2)\in\mathbb{R}_+^2$ to get small quantities on the right hand side.

Proof of Equation (A.1)*:* The first step consists in working conditionally on $\mathcal{F}$, describing this probability as a product of conditional probabilities close to one, as well as in deriving a Poisson approximation. To this aim, we define for $m\in\mathbb{N}$:

$$\theta^{(12)}(m) := \mathbf{1}_{\{\tau_m^K\le\tilde{T}_\varepsilon^K\}}\mathbf{1}_{\{\tilde{N}_a(\tau_m^K)-\tilde{N}_a(\tau_{m-1}^K)=1\}}(p_{aA}^{(1)}+p_{aA}^{(2)})(\tilde{N}_A,\tilde{N}_a)(\tau_{m-1}^K),$$

where we recall the definition of the $p^{(i)}_{\alpha\alpha'}$ in Definition 6.1. Notice that Remark 1 p. 21 implies that for $\rho\in\{1,2\}$, $n_A\in I_\varepsilon^K\pm 1$ and $l<\lfloor\varepsilon K\rfloor$,
(A.5)

$$(1-c\varepsilon)(r_1+r_2)^\rho\Big(\sum_{k=1}^{l-1}\frac{\mathbb{E}^{(1)}_{(n_A,l)}U_k^K(1)}{(k+1)^\rho}\Big) \le \mathbb{E}^{(1)}_{(n_A,l)}\Big[\sum_{m=1}^{\infty}(\theta^{(12)}(m)\mathbf{1}_{\{\tilde{N}_a(\tau_m^K)<l\}})^\rho\Big] \le (r_1+r_2)^\rho\Big(\sum_{k=1}^{l-1}\frac{\mathbb{E}^{(1)}_{(n_A,l)}U_k^K(1)}{(k+1)^\rho}\Big).$$

Then, similarly as in [18], we have for $n_A\in I_\varepsilon^K\pm 1$ and $l<\lfloor\varepsilon K\rfloor$

$$\mathbb{P}^{(1)}_{(n_A,l)}(NR(i)^{(1)}|\mathcal{F}) = \prod_{m=1}^{\infty}(1-\theta^{(12)}(m)), \quad \mathbb{P}^{(1)}_{(n_A,l)} - \text{a.s.}$$

If we introduce the variable,

$$\eta^{(12)} := \sum_{m=1}^{\infty}\theta^{(12)}(m),$$

which will play the role of X_1 in (A.4), we get by following the path of Lemma 3.6 in [18]:

$$\text{(A.6)} \qquad \mathbb{E}^{(1)}_{(n_A,l)}\Big|\prod_{m=1}^{\infty}(1-\theta^{(12)}(m)) - \exp(-\eta^{(12)})\Big| \le \mathbb{E}^{(1)}_{(n_A,l)}\Big[\sum_{m=1}^{\infty}(\theta^{(12)}(m))^2\Big] \le \frac{c}{\log^2 K},$$

for K large enough, $n_A\in I_\varepsilon^K\pm 1, l<\lfloor\varepsilon K\rfloor$ and a finite c (which can be chosen independently of l), where we used Equations (5.10) (5.11) and (A.5), and Condition (1.1) for the last inequality. Next we introduce

an approximation of the random variable $\eta^{(12)}$, namely

$$\tilde{\eta}^{(12)} := \sum_{m=1}^{\infty} \theta^{(12)}(m) \mathbf{1}_{\{\tilde{N}_a(\tau_m^K) \geq \tilde{N}_a(0)\}}, \tag{A.7}$$

which will play the role of X_2 in (A.4). For $n_A \in I_\varepsilon^K \pm 1$ and $l \leq \lfloor \varepsilon K \rfloor$:

$$0 \leq \mathbb{E}^{(1)}_{(n_A,l)}[\eta^{(12)} - \tilde{\eta}^{(12)}] = \mathbb{E}^{(1)}_{(n_A,l)}\Big[\sum_{m=1}^{\infty} \theta^{(12)}(m) \mathbf{1}_{\{\tilde{N}_a(\tau_m^K) < l\}}\Big] \tag{A.8}$$

$$\leq \frac{r_1+r_2}{s_+(\varepsilon)s_-^2(\varepsilon)} \sum_{k=1}^{l-1} \frac{(1-s_-(\varepsilon))^{l-k}}{k+1} \leq c\frac{(r_1+r_2)}{l}, \tag{A.9}$$

for a finite c and ε small enough, where we used (A.5) and (5.11) for the first inequality, and (B.2) for the second one. This latter ensures that c can be chosen independently of l. The expected value of $\tilde{\eta}^{(12)}$ can be bounded by using (A.5), (5.10) and (B.2)

$$\begin{aligned} \mathbb{E}^{(1)}_{(n_A,l)}[\tilde{\eta}^{(12)}] &\geq (1-c\varepsilon)(r_1+r_2) \sum_{k=l}^{\lfloor \varepsilon K \rfloor -1} \frac{1}{k+1}\Big(\frac{1-(1-s)^{\lfloor \varepsilon K \rfloor - k} - (1-s)^{k+1}}{s} - c\varepsilon\Big) \\ &\geq (1-c\varepsilon)\frac{r_1+r_2}{s} \log \frac{\lfloor \varepsilon K \rfloor}{l} - \frac{c}{\log K}, \end{aligned} \tag{A.10}$$

for a finite c and ε small enough. For the upper bound we get similarly,

$$\mathbb{E}^{(1)}_{(n_A,l)}[\tilde{\eta}^{(12)}] \leq (1+c\varepsilon)\frac{r_1+r_2}{s} \log \frac{\lfloor \varepsilon K \rfloor}{l}. \tag{A.11}$$

The last step consists in bounding the variance of $\tilde{\eta}^{(12)}$. As the calculation of this variance is quite involved, we introduce an approximation of $\tilde{\eta}^{(12)}$, namely

$$\tilde{\tilde{\eta}}^{(12)} := \sum_{m=1}^{\infty} \mathbf{1}_{\{\tilde{N}_a(\tau_{m-1}^K) \geq \tilde{N}_a(0)\}} \mathbf{1}_{\{\tilde{N}_a(\tau_m^K) - \tilde{N}_a(\tau_{m-1}^K) = 1\}} \frac{r_1+r_2}{\tilde{N}_a(\tau_{m-1}^K)+1} = \sum_{k=\tilde{N}_a(0)}^{\lfloor \varepsilon K \rfloor -1} \frac{r_1+r_2}{k+1} U_k^K(1).$$

Equation (6.2) yields $(1-c\varepsilon)\tilde{\tilde{\eta}}^{(12)} \leq \tilde{\eta}^{(12)} \leq \tilde{\tilde{\eta}}^{(12)}$ for a finite c and ε small enough. Hence

$$\begin{aligned} \Big|\mathrm{Var}^{(1)}_{(n_A,l)} \tilde{\eta}^{(12)} - \mathrm{Var}^{(1)}_{(n_A,l)} \tilde{\tilde{\eta}}^{(12)}\Big| &\leq c\varepsilon \mathbb{E}^{(1)}_{(n_A,l)}\Big[\Big(\tilde{\tilde{\eta}}^{(12)}\Big)^2\Big] \\ &\leq c\varepsilon(r_1+r_2)^2 \sum_{k,k'=l}^{\lfloor \varepsilon K \rfloor -1} \frac{\mathbb{E}^{(1)}[(U_k^K(1))^2] + \mathbb{E}^{(1)}[(U_{k'}^K(1))^2]}{(k+1)(k'+1)} \leq c\varepsilon, \end{aligned} \tag{A.12}$$

where we used (5.15) and (B.3) which ensure that $U_k^K(1)$ is smaller than a geometric random variable with parameter $q_k^{(s_-(\varepsilon),s_+(\varepsilon))} \geq s_-(\varepsilon)$. Thus it is enough to bound $\mathrm{Var}^{(1)}_{(n_A,l)} \tilde{\tilde{\eta}}^{(12)}$. Thanks to (5.12) and Condition (1.1) we get:

$$\begin{aligned} \mathrm{Var}^{(1)}_{(n_A,l)} \tilde{\tilde{\eta}}^{(12)} &= (r_1+r_2)^2 \sum_{k,k'=l}^{\lfloor \varepsilon K \rfloor -1} \frac{\mathrm{Cov}^{(1)}_{(n_A,l)}(U_k^K(1), U_{k'}^K(1))}{(k+1)(k'+1)} \\ &\leq 2(r_1+r_2)^2 \sum_{k \leq k'=l}^{\lfloor \varepsilon K \rfloor -1} \frac{\lambda_\varepsilon^{(k'-k)/2} + \varepsilon}{(k+1)(k'+1)} \leq c\frac{\log \lfloor \varepsilon K \rfloor}{\log^2 K}(c + \varepsilon \log \lfloor \varepsilon K \rfloor). \end{aligned}$$

Recalling (A.12) and again Condition (1.1), we finally obtain

$$\limsup_{K \to \infty} \mathrm{Var}^{(1)}_{(n_A,l)} \tilde{\eta}^{(12)} \leq c\varepsilon, \tag{A.13}$$

for a finite c independent of l and ε small enough. Applying (A.4) with $X_1 = \eta^{(12)}$ and $X_2 = \tilde{\eta}^{(12)}$ yields

$$\Big|\mathbb{P}^{(1)}_{(n_A,l)}(NR(i)^{(1)}) - e^{-\frac{r_1+r_2}{s}\log\frac{\lfloor \varepsilon K\rfloor}{l}}\Big| \leq \mathbb{E}^{(1)}_{(n_A,l)}\Big|\prod_{m=1}^{\infty}(1-\theta^{(12)}(m)) - \exp(-\eta^{(12)})\Big|$$
$$+\mathbb{E}^{(1)}_{(n_A,l)}[\eta^{(12)}-\tilde{\eta}^{(12)}] + \sqrt{\mathrm{Var}^{(1)}_{(n_A,l)}\tilde{\eta}^{(12)}} + \Big|\mathbb{E}^{(1)}_{(n_A,l)}[\tilde{\eta}^{(12)}] - \frac{r_1+r_2}{s}\log\frac{\lfloor\varepsilon K\rfloor}{l}\Big|.$$

We end the proof of Equation (A.1) with Inequalities (A.6), (A.8), (A.13), (A.10) and (A.11).

Proof of (A.2)*:* There is a supplementary difficulty due to the randomness of $\tilde{N}_a(\tau^K_{R(i,2)})$. In the previous case we were interested in an event before the first hitting of $\lfloor\varepsilon K\rfloor$, while in the current case, the conditioning on the value of $\tilde{N}_a(\tau^K_{R(i,2)})$ does not tell us how many times $\tilde{N}_a$ has hit this value before. This is why we have introduced $NR(l,\sigma,i)$ and $NR(l,\zeta,i)$ in (6.13). Define for $m \geq 1$,

$$\theta^{(1)}(m) := \mathbf{1}_{\{\tau^K_m \leq \tilde{T}^K_\varepsilon\}}\mathbf{1}_{\{\tilde{N}_a(\tau^K_m)-\tilde{N}_a(\tau^K_{m-1})=1\}}p^{(1)}_{aA}((\tilde{N}_A,\tilde{N}_a)(\tau^K_m)).$$

We again condition on the trait population process and get for $n_A \in I^K_\varepsilon \pm 1$ and $k \leq l < \lfloor\varepsilon K\rfloor$,

$$\mathbb{P}^{(1)}_{(n_A,k)}(NR(l,\sigma,i)|\mathcal{F}) = \prod_{m=1}^{\sigma^K_l(1)}(1-\theta^{(1)}(m)), \quad \mathbb{P}^{(1)}_{(n_A,k)} - \text{a.s.}, \tag{A.14}$$

and the same expression with σ replacing ζ. We define the corresponding parameters for the Poisson approximation as follows:

$$\eta^{(1),-}_l := \sum_{m=1}^{\sigma^K_l(1)}\theta^{(1)}(m), \quad \text{and} \quad \eta^{(1),+}_l := \sum_{m=1}^{\zeta^K_l(1)}\theta^{(1)}(m).$$

They will play the role of X_1 in (A.4). We will show that both can be approximated by:

$$\tilde{\eta}^{(1)}_l := \sum_{m=1}^{\zeta^K_l(1)}\theta^{(1)}(m)\mathbf{1}_{\{\tilde{N}_a(0)\leq\tilde{N}_a(\tau^K_m)\leq l\}}, \tag{A.15}$$

which will play the role of X_2 in (A.4). Recall Definitions (5.8), (5.17) and (5.20). On the one hand, for $n_A \in I^K_\varepsilon \pm 1$ and $k < \lfloor\varepsilon K\rfloor$,

$$\mathbb{E}^{(1)}_{(n_A,k)}[\eta^{(1),+}_l - \tilde{\eta}^{(1)}_l] = \mathbb{E}^{(1)}_{(n_A,k)}\Big[\sum_{m=1}^{\zeta^K_l(1)}\theta^{(1)}(m)(\mathbf{1}_{\{N^K_a(\tau^K_m)<k\}} + \mathbf{1}_{\{N^K_a(\tau^K_m)>l\}})\Big]$$
$$\leq \mathbb{E}^{(1)}[D^K_k(1)]\sum_{j=1}^{k-1}\sup_{n_A\in I^K_\varepsilon}p^{(1)}_{aA}(n_A,j)\sup_{n_A\in I^K_\varepsilon\pm1}\mathbb{E}^{(1)}_{(n_A,k-1)}[U^K_{n_A,k,j}(1)]$$
$$+\mathbb{E}^{(1)}[U^K_l(1)]\sum_{j=l+1}^{\lfloor\varepsilon K\rfloor}\sup_{n_A\in I^K_\varepsilon}p^{(1)}_{aA}(n_A,j)\sup_{n_A\in I^K_\varepsilon\pm1}\mathbb{E}^{(1)}_{(n_A,l+1)}[U^K_{n_A,l,j}(1)|\sigma^K_l(1)<\infty], \tag{A.16}$$

where we used that in the first phase, under $\mathbb{P}^{(1)}$, the number of excursions below k (resp. above l) is equal to $D^K_k(1)$ (resp. $U^K_l(1)-1$). Applying Inequality (5.10), Lemma 5.3, and Equation (6.2), we get the existence of a finite c such that for ε small enough:

$$\mathbb{E}^{(1)}_{(n_A,k)}[\eta^{(1),+}_l - \tilde{\eta}^{(1)}_l] \leq cr_1\sum_{j=1}^{\lfloor\varepsilon K\rfloor}\frac{\mu_\varepsilon^{|j-l|}}{j+1} \leq \frac{c}{\log K},$$

as $\mu_\varepsilon \in (0,1)$ for ε small enough and by Condition (1.1). On the other hand, by using the same results as in (A.16), we get

$$\begin{aligned}\mathbb{E}^{(1)}_{(n_A,k)}[|\eta_l^{(1),-}-\tilde{\eta}_l^{(1)}|] &\leq \mathbb{E}^{(1)}_{(n_A,k)}\Big[\sum_{m=1}^{\sigma_l^K(1)}\theta^{(1)}(m)\mathbf{1}_{\{\tilde{N}_a(\tau_m^K)<k\}}+\sum_{m=\sigma_l^K(1)+1}^{\zeta_l^K(1)}\theta^{(1)}(m)\mathbf{1}_{\{k\leq\tilde{N}_a(\tau_m^K)\leq l\}}\Big]\\ &\leq cr_1\Big(\sum_{j=1}^{k-1}\frac{\mu_\varepsilon^{k-j}}{j+1}+\sum_{j=k}^{l-1}\frac{\mu_\varepsilon^{l-j}}{j+1}\Big)\leq\frac{c}{\log K}.\end{aligned}$$

This shows that it is sufficient to use $\tilde{\eta}_l^{(1)}$ for the Poisson approximation. From (A.6) we deduce that this approximation holds true up to terms of order $1/\log^2 K$. Recalling once again (A.4), we see that it only remains to calculate the expected value of $\tilde{\eta}_l^{(1)}$ and to bound its variance. The expectation can be approximated in the same way as the expected value of $\tilde{\eta}_l^{(12)}$ from the previous part in (A.10) and (A.11):

$$(1-c\varepsilon)\frac{r_1}{s}\log\frac{l-1}{k}-\frac{c}{\log K}\leq\mathbb{E}^{(1)}_{(n_A,k)}[\tilde{\eta}_l^{(1)}]\leq(1+c\varepsilon)\frac{r_1}{s}\log\frac{l-1}{k}. \tag{A.17}$$

A comparison of the definitions of $\tilde{\eta}_l^{(1)}$ in (A.15) and $\tilde{\eta}^{(12)}$ in (A.7) shows that the variance of $\tilde{\eta}_l^{(1)}$ can be bounded by the same expression, that is, a constant times ε. This ends the proof of Equation (A.2).

Proof of Equation (A.3) It can be done in a similar way as for Equations (A.1) and (A.2). We have the following lower and upper bounds:

$$\prod_{m=1}^{\zeta_l^K(1)}\Big[1-p_{AA}^{(2)}(\tilde{N}_A,\tilde{N}_a)(\tau_m^K)\Big]\ \leq\ 1-\mathbb{P}^{(1)}(RA(l,i)|\mathcal{F})\ \leq\ \prod_{m=1}^{\sigma_l^K(1)}\Big[1-p_{AA}^{(2)}(\tilde{N}_A,\tilde{N}_a)(\tau_m^K)\Big]. \tag{A.18}$$

Once again we aim at deriving a Poisson approximation. As a birth event in the A-population is needed to see a recombination within the A-population, bounds on the expected number of jumps will concern the process $\tilde{N}_A$ and we have to use Lemma 5.4. □

Appendix B. Technical results

This section is dedicated to technical results needed in the proofs. First we recall a well known result on the hitting times of birth and death processes which can be found in [18] Lemma 3.1:

Proposition 7. *Let $Z=(Z_t)_{t\geq 0}$ be a birth and death process with individual birth and death rates b and d. For $i\in\mathbb{Z}^+$, $T_i=\inf\{t\geq 0, Z_t=i\}$ and $\mathbb{P}_i$ is the law of Z when $Z_0=i$. Then for $(i,j,k)\in\mathbb{Z}_+^3$ such that $j\in(i,k)$,*

$$\mathbb{P}_j(T_k<T_i)=\frac{1-(d/b)^{j-i}}{1-(d/b)^{k-i}}. \tag{B.1}$$

We also recall Lemma 3.5 in [18] and the first part of Equation (A.16) in [19] which are used several times:

Lemma B.1.

- *If $a>1$ there is a C such that for every $N\in\mathbb{N}$,*

$$\sum_{j=1}^{N}\frac{a^j}{j}\leq\frac{Ca^N}{N}. \tag{B.2}$$

- *Recall Definition* (5.14)*. Then for $(s_1,s_2)\in(0,1)^2$ and $k<\lfloor\varepsilon K\rfloor$,*

$$q_k^{(s_1\wedge s_2,s_1\vee s_2)}\geq s_1\wedge s_2. \tag{B.3}$$

Finally, we state two technical results. The first one can be proven by using characteristic functions, the proof of the second Lemma is given below:

Lemma B.2. *Let V be a geometric random variable with parameter p_1 and $(G^i, i \in \mathbb{N})$ a sequence of independent geometric random variables with parameter p_2, independent of V. Then the random variable:*

$$Z := \sum_{i \le V} G^i$$

is geometrically distributed with parameter $p_1 p_2$.

Lemma B.3. *Let $(c_N, N \in \mathbb{N})$ be a bounded sequence of $\mathbb{R}$. Then there exists a finite constant c such that*

$$\limsup_{N\to\infty} \sup_{k \le N} \Big| \sum_{l=1}^{k-1} \frac{e^{\frac{c_N}{\log N}\log l}}{l+1} - \frac{\log N}{c_N}(e^{\frac{c_N}{\log N}\log k} - 1) \Big| \le c.$$

Proof. We prove the Lemma for a sequence $(c_N, N \in \mathbb{N})$ in $\mathbb{R}^*$ and extend the result by using the convention

$$\Big(\frac{\log N}{c_N}(e^{\frac{c_N}{\log N}\log k} - 1)\Big)_{|c_N=0} = \log k.$$

The idea is to compare the sum with the integral

$$\int_1^k x^{\frac{c_N}{\log N}-1} dx = \frac{\log N}{c_N}(e^{\frac{c_N}{\log N}\log k} - 1).$$

Let l be in $\{1, ..., N-1\}$. Then we have

$$\begin{aligned} \int_l^{l+1} x^{\frac{c_N}{\log N}-1} dx - \frac{l^{\frac{c_N}{\log N}}}{l+1} &= \frac{\log N}{c_N}\Big((l+1)^{\frac{c_N}{\log N}} - l^{\frac{c_N}{\log N}} - \frac{c_N}{\log N}\frac{l^{\frac{c_N}{\log N}}}{l+1}\Big) \\ &= \frac{\log N}{c_N} l^{\frac{c_N}{\log N}} \Big(\Big(1+\frac{1}{l}\Big)^{\frac{c_N}{\log N}} - 1 - \frac{c_N}{(l+1)\log N}\Big). \end{aligned}$$

An application of the Taylor-Lagrange formula yields that

$$\Big(1+\frac{1}{l}\Big)^{\frac{c_N}{\log N}} - 1 = \frac{c_N}{l \log N} + \frac{c_N}{\log N}\Big(\frac{c_N}{\log N} - 1\Big)\frac{1}{2l^2}\Big(1+x\Big)^{\frac{c_N}{\log N}-2}$$

where x belongs to $[0, 1/l]$. As the sequence $(c_N, N \in \mathbb{N})$ is bounded, we deduce that there exists a finite constant c such that

$$\Big| \int_l^{l+1} x^{\frac{c_N}{\log N}-1} dx - \frac{l^{\frac{c_N}{\log N}}}{l+1} \Big| \le \frac{c}{l^2}.$$

This ends up the proof of Lemma B.3. □

Acknowledgements: *The authors would like to thank Jean-François Delmas, Sylvie Méléard and Anja Sturm for their careful reading of this paper. They also want to thank an anonymous reviewer for several suggestions and improvements. This work was partially funded by project MANEGE 'Modèles Aléatoires en Écologie, Génétique et Évolution' of the French national research agency ANR-09-BLAN-0215, Chair Modélisation Mathématique et Biodiversité Veolia Environnement- Ecole Polytechnique-Museum National d'Histoire Naturelle-Fondation X and the French national research agency ANR-11-BSV7-013-03, the DFG through SPP priority programme 1590 and the RTG 1644, 'Scaling Problems in Statistics'.*

References

[1] N. H. Barton. The effect of hitch-hiking on neutral genealogies. Genetical Research, 72(2):123–133, 1998.

[2] R. Brink-Spalink. Stochastic Models in Population Genetics: The Impact of Selection and Recombination. PhD thesis, Georg-August University, 2014.

[3] N. Champagnat. A microscopic interpretation for adaptive dynamics trait substitution sequence models. Stochastic Processes and their Applications, 116(8):1127–1160, 2006.

[4] N. Champagnat, P.-E. Jabin, and S. Méléard. Adaptation in a stochastic multi-resources chemostat model. Journal de Mathématiques Pures et Appliquées, 101(6):755–788, 2014.

[5] N. Champagnat and S. Méléard. Polymorphic evolution sequence and evolutionary branching. Probability Theory and Related Fields, 151(1-2):45–94, 2011.

[6] P. Collet, S. Méléard, and J. A. Metz. A rigorous model study of the adaptive dynamics of mendelian diploids. Journal of Mathematical Biology, pages 1–39, 2011.

[7] C. Coron. Slow-fast stochastic diffusion dynamics and quasi-stationarity for diploid populations with varying size. Journal of Mathematical Biology, 72(1):171–202, 2015.

[8] M. A. Diniz and A. Polpo. A simple proof for the multinomial version of the representation theorem. In The Contribution of Young Researchers to Bayesian Statistics, pages 15–18. Springer, 2014.
[9] R. Durrett and J. Schweinsberg. Approximating selective sweeps. Theoretical population biology, 66(2):129–138, 2004.
[10] A. Etheridge, P. Pfaffelhuber, and A. Wakolbinger. An approximate sampling formula under genetic hitchhiking. The Annals of Applied Probability, 16(2):685–729, 2006.
[11] S. Ethier and T. Kurtz. Markov processes: Characterization and convergence, 1986, 1986.
[12] N. Fournier and S. Méléard. A microscopic probabilistic description of a locally regulated population and macroscopic approximations. The Annals of Applied Probability, 14(4):1880–1919, 2004.
[13] N. L. Kaplan, R. Hudson, and C. Langley. The" hitchhiking effect" revisited. Genetics, 123(4):887–899, 1989.
[14] S. Leocard. Selective sweep and the size of the hitchhiking set. Advances in Applied Probability, 41(3):731–764, 2009.
[15] J. A. Metz, S. A. Geritz, G. Meszéna, F. J. Jacobs, and J. Van Heerwaarden. Adaptive dynamics, a geometrical study of the consequences of nearly faithful reproduction. Stochastic and spatial structures of dynamical systems, 45:183–231, 1996.
[16] T. Ohta and M. Kimura. The effect of selected linked locus on heterozygosity of neutral alleles (the hitch-hiking effect). Genetical research, 25(03):313–325, 1975.
[17] P. Pfaffelhuber and A. Studeny. Approximating genealogies for partially linked neutral loci under a selective sweep. Journal of mathematical biology, 55(3):299–330, 2007.
[18] J. Schweinsberg and R. Durrett. Random partitions approximating the coalescence of lineages during a selective sweep. The Annals of Applied Probability, 15(3):1591–1651, 2005.
[19] C. Smadi. An eco-evolutionary approach of adaptation and recombination in a large population of varying size. Stochastic Processes and their Applications, 125(5):2054–2095, 2015.
[20] J. M. Smith and J. Haigh. The hitch-hiking effect of a favourable gene. Genet Res, 23(1):23–35, 1974.
[21] W. Stephan, Y. S. Song, and C. H. Langley. The hitchhiking effect on linkage disequilibrium between linked neutral loci. Genetics, 172(4):2647–2663, 2006.
[22] W. Stephan, T. H. Wiehe, and M. W. Lenz. The effect of strongly selected substitutions on neutral polymorphism: analytical results based on diffusion theory. Theoretical Population Biology, 41(2):237–254, 1992.